AD-A047 277 AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OHIO SCHO--ETC F/G 5/1
A MODEL FOR FORECASTING THE COMPUTER REQUIREMENTS FOR THE AIR F--ETC(U)
SEP 77 H H DETJEN, D C JOHNSTON

UNCLASSIFIED AFIT-LSSR-1-77B NL

1 OF 2
AD
A047 277

(3)

A MODEL FOR FORECASTING THE COMPUTER REQUIREMENTS FOR THE AIR FORCE INSTITUTE OF TECHNOLOGY SCHOOL OF SYSTEMS AND LOGISTICS

Herman H. Detjen, Sqn Ldr, RAAF
David C. Johnston, 2nd Lt, USAF

LSSR 1-77B

The contents of the document are technically accurate, and no sensitive items, detrimental ideas, or deliterious information are contained therein. Furthermore, the views expressed in the document are those of the author and do not necessarily reflect the views of the School of Systems and Logistics, the Air University, the United States Air Force, or the Department of Defense.

USAF SCN 75-20B AFIT Control Number LSSR 1-77B

AFIT RESEARCH ASSESSMENT

The purpose of this questionnaire is to determine the potential for current and future applications of AFIT thesis research. Please return completed questionnaires to: AFIT/SLGR (Thesis Feedback), Wright-Patterson AFB, Ohio 45433.

1. Did this research contribute to a current Air Force project?

a. Yes b. No

2. Do you believe this research topic is significant enough that it would have been researched (or contracted) by your organization or another agency if AFIT had not researched it?

a. Yes b. No

3. The benefits of AFIT research can often be expressed by the equivalent value that your agency received by virtue of AFIT performing the research. Can you estimate what this research would have cost if it had been accomplished under contract or if it had been done in-house in terms of manpower and/or dollars?

a. Man-years ________ $ ________ (Contract).

b. Man-years ________ $ ________ (In-house).

4. Often it is not possible to attach equivalent dollar values to research, although the results of the research may, in fact, be important. Whether or not you were able to establish an equivalent value for this research (3 above), what is your estimate of its significance?

a. Highly Significant b. Significant c. Slightly Significant d. Of No Significance

5. Comments:

________________________ ________________________
Name and Grade Position

________________________ ________________________
Organization Location

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

14

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
AFIT-LSSR-1-77B ✓

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

9

4. TITLE (and Subtitle)

6 A MODEL FOR FORECASTING THE COMPUTER REQUIREMENTS FOR THE AIR FORCE INSTITUTE OF TECHNOLOGY SCHOOL OF SYSTEMS AND LOGISTICS,

5. TYPE OF REPORT & PERIOD COVERED
Master's Thesis,

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)

10 Herman H. Detjen Squadron Leader, RAAF
David C. Johnston Second Lieutenant, USAF

8. CONTRACT OR GRANT NUMBER(s)

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Graduate Education Division ✓
School of Systems and Logistics
Air Force Institute of Technology, WPAFB OH

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
Department of Research and Administrative Management (LSGR)
AFIT/LSGR, WPAFB OH 45433

12. REPORT DATE
11 Sep[redacted]77

13. NUMBER OF PAGES
116 12 134p.

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

APPROVED FOR PUBLIC RELEASE AFR 190-17.

JERRAL F. GUESS, CAPT, USAF
Director of Information

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Computer System Usage; Time-Series Analysis; Regression Analysis; Computer Utilization; Forecasting Models

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Thesis Chairman: Eugene E. Jones, Lieutenant Colonel, USAF

see back

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

012250

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

This thesis analyzes and forecasts the usage pattern of the Air Force Logistics Command's CREATE computer system by the Air Force Institute of Technology School of Systems and Logistics. The analysis uses exponential smoothing in time-series forecasting, autocorrelation, bivariate correlation, and linear regression techniques to determine relationships among monthly aggregated computer usage parameters. The monthly aggregated computer parameters included batch jobs, CPU hours batch, CPU hours time-sharing, core hours, log-on hours, log-ons and lines CARDIN. While time-series forecasting can be satisfactorily used in most cases for forecasting purposes, autocorrelation analysis provided insight into the cyclic relationships between the data and time. The authors concluded that an unnecessarily long time-series can introduce inaccuracies into a forecast because the process of use of the system changes over time. Bivariate correlation analysis and linear regression demonstrated both simple and multiple linear relationships between the variables used. The authors concluded that these linear relationships could be used in defining the requirements for a new computing system to replace an existing system or part of a system.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

LSSR 1-77B

A MODEL FOR FORECASTING THE COMPUTER REQUIREMENTS FOR THE AIR FORCE INSTITUTE OF TECHNOLOGY SCHOOL OF SYSTEMS AND LOGISTICS

A Thesis

Presented to the Faculty of the School of Systems and Logistics

of the Air Force Institute of Technology

Air University

In Partial Fulfillment of the Requirements for the

Degrees of Master of Science in Logistics Management

and Master of Science in Facilities Management

By

Herman H. Detjen, ARMIT(CommEng)
Squadron Leader, RAAF

David C. Johnston, BS
Second Lieutenant, USAF

September 1977

Approved for public release;
distribution unlimited

This thesis, written by

Squadron Leader Herman H. Detjen

and

Second Lieutenant David C. Johnston

has been accepted by the undersigned on behalf of the faculty of the School of Systems and Logistics in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN LOGISTICS MANAGEMENT
(Squadron Leader Herman H. Detjen)

MASTER OF SCIENCE IN FACILITIES MANAGEMENT
(Second Lieutenant David C. Johnston)

DATE: 7 September 1977

COMMITTEE CHAIRMAN

ACKNOWLEDGEMENTS

First, we wish to thank our wives, Pauline and Mitzi, for their patience, understanding and support during the long hours we spent on this research. Without their help, we would have had difficulty in completing this effort.

We wish to express our thanks to Cliff Brown (AFLC/ACDSM CREATE Management) for the assistance he rendered in providing the data for this research. His co-operation simplified our data collection tasks considerably.

We also wish to thank Richard Sowar for the extensive amount of graphics support he provided and Major James Abbott and the AFIT/ADS staff for the use of their computer terminal.

Last, but not least, we wish to thank our adviser, Lt Col Eugene E. Jones for the many hours of guidance and assistance he has provided. His knowledge of forecasting techniques was of immeasurable assistance to us in completing this research.

TABLE OF CONTENTS

CHAPTER	Page

CHAPTER Page

APPENDIXES Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE		Page

TABLE		Page
28.	PEARSON'S BIVARIATE CORRELATION COEFFICIENTS AND LEVEL OF SIGNIFICENCE --CREATE-- APRIL 1976 TO MARCH 1977	111
29.	MULTIPLE LINEAR REGRESSION RESULTS--CREATE . . .	112
30.	TCAST ANALYSIS--CREATE--APRIL 1975 TO MARCH 1977--LEAD TIME= 12	113

CHAPTER I

INTRODUCTION

The Air Force Institute of Technology (AFIT), located on Wright-Patterson Air Force Base (WPAFB), Ohio, performs its assigned mission through the educational and training programs of three separate schools: the School of Engineering (AFIT/EN), the School of Civil Engineering (AFIT/DE), and the School of Systems and Logistics (AFIT/LS).

The computer support for both administrative and scientific requirements of the three schools is provided by two separate systems. The first is the Aeronautical System Division's (ASD) Control Data Corporation (CDC) Dual 6600 computing system. The second is the Air Force Logistics Command's (AFLC) Honeywell-635 Computational Resources for Engineering And Simulation, Training and Education (CREATE) computing system. AFIT/EN and AFIT/DE have access to both computing systems, while AFIT/LS is connected only to the CREATE system.

Statement of the Problem

At the present time, neither the CDC nor the CREATE system has a method for forecasting computer support requirements based on the number of log-ons, computer

processor time and other computer operating parameters. To allow for assessment of the impact of predicted growth of required computing services through changes in AFIT (i.e. curriculum, student population and like factors), a forecasting technique is needed (5).

Scope

This research will provide an analysis of the CREATE resources used only by AFIT/LS. No analysis of the CDC 6600 system used by AFIT/EN or AFIT/DE, nor that portion of CREATE used by these two schools is intended. However, an assessment of the research methodology for application to other computing systems should be possible as a part of the research.

Background

School of Systems and Logistics

AFIT/LS offers a twelve-month resident graduate education program leading to a Master of Science in Logistics Management or Facilities Management. The school also conducts a continuing education program which provides short courses in systems and logistics as required to meet the needs of the Department of Defense (9:96-136). In pursuing these programs, AFIT/LS uses the computing facilities provided by AFLC.

The AFLC/CREATE System

The present AFLC/CREATE system (hereafter referred to as CREATE) is a Honeywell 635 (see Figure 1). The Honeywell 635 consists of two processors, four memory modules (64000 36-bit words each) and two input-output control (IOC) units. The immediate access storage units are disc drives. The communication unit between the users and the computer is a DATANET 355. Over 120 time-sharing terminals and batch-remote units of varying speeds are connected to the computer from WPAFB and seven remote user locations. Approximately 60 of the time-sharing terminals and 4 batch-remote units are located at WPAFB (8).

The workload for CREATE is required to meet one of the following applications: (1) engineering computation, (2) logistics research, or (3) education (AFIT only). In addition, any application must meet at least one specific workload criteria. The criteria are identified as statistical analysis, computational methods, simulation, computer aided instructions, computer aided design, mathematical programming or education (11:5).

The AFIT/LS CREATE System

The AFIT/LS computer facility (hereafter referred to as AFIT/CREATE) is a portion of the total CREATE system. AFIT/CREATE consists of approximately seventeen time-sharing terminals and one Honeywell 115 remote batch unit which provides printer, punch and card reader capability.

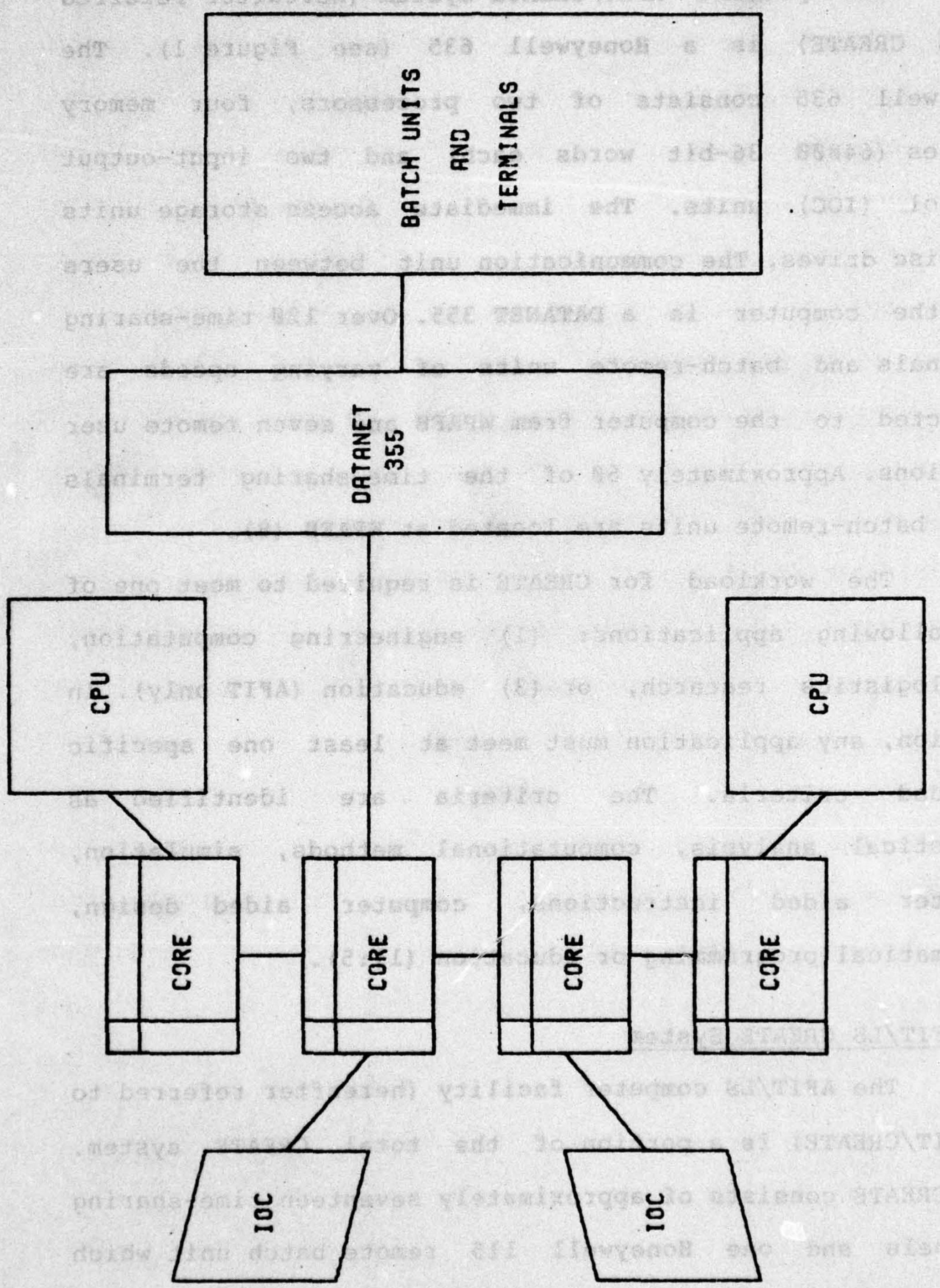


Figure 1. HONEYWELL 365 COMPUTING SYSTEM

Forecasting Model Development

Relatively little research was found that provided information on forecasting techniques used for predicting system utilization. Most research seems to be directed toward analyzing computer system performance rather than forecasting future demands on a system.

Anderson and Purnell provided the initial effort in developing a forecasting model for the AFIT/CREATE system. Their model applied linear regression methods and time-series forecasting techniques to monthly-aggregated computer usage data. It was found that time-series forecasting was more appropriate. However, a longer time data base which could incorporate two complete cycles of Graduate Logistics "A" and "B" classes without schedule interruptions was required. This would "allow increased use of time-series forecasting techniques and the elimination of linear regression methods [1:67]."

Anderson and Purnell concluded that the time-series model was capable of incorporating changes in the data base which are known to be programmed for the forecast period. The time-series model also provided a low error term for each computer operating parameter used (1:66).

However, Anderson and Purnell also showed that multiple linear regression of aggregated computer operating parameters provided an efficient method of defining the association between the operating parameters (1:31-39). The

usefulness of regression techniques for forecasting compared to time-series analysis was not fully evaluated. They indicated the consistency of the data base was lacking for time-series forecasting techniques to provide as accurate a forecast as necessary. Yet the evaluation of regression techniques to overcome these types of problems was not addressed.

Hunt, Diehr and Garnatz undertook similar studies on the use of the University of Washington computer system. Their studies were directed toward determining the influence of different categories of users (students, faculty, etc.) on the computing system and the methods of cost accounting computer usage. Descriptive statistics and correlation analysis were applied to model computer use. The most satisfactory results were obtained using less conventional cluster analysis techniques (4:232-238).

Hunt, et al. maintained that a knowledge of user characteristics is important in planning the use of any large computing system. Using descriptive statistics, they concluded: (a) mean values of computer usage parameters for individual jobs were not good descriptors of the population, and (b) the distributions of the frequency plots of the computer usage parameters were positively skewed with means in regions of very low density. That is, the "average" user is not the typical user (4:238).

Within the group of users of a computing system, there are categories of users who use the system in a particular manner. Using cluster analysis, they showed that central tendency measures were reasonable descriptors for the "average" user in each distinct category. They concluded that cluster analysis techniques would ". . . aid in identifying the characteristics of existential, rather than postulated computer users [4:238]."

Each of the research approaches described above provides a different technique for identifying the characteristics of computer usage and applying this knowledge for forecasting and planning purposes. This research effort will apply the principles used by Anderson and Purnell in developing a suitable forecasting model.

Justification

Watson (12) stated that most Air Force computing installations collect and record accounting data. However, seldom is any use made of these data except to "charge" for computing services. Accounting data record the quantities of system resources used. As such, these data are suitable for evaluating workload characteristics. In particular, compiling trends of past usage and performance is useful in forecasting future demands on the system and provides more effective management of computing systems.

The importance of evaluation of computer utilization has been recognized for some time. An installation needs to identify specific causes of saturation so that limited resources may be better managed to improve system performance (6:671). In a computing system such as the AFIT/CREATE system, there is generally a steady increase in the number of users and use of system resources. The characteristics of these increases and the effect on computing resources must be known to be able to anticipate bottlenecks and to plan for orderly expansion (4:231).

Demands on the CREATE system had continued to increase and had caused a decline in the ability to respond to user demands (10). If these demands could be forecast, it would allow better management of available resources. Also, planning of expansion to meet growth requirements would be possible before saturation of available resources and deterioration of services occurs.

Neither the CREATE system nor the AFIT/CREATE system has been analyzed for growth and projected workloads with conclusive results. Such an analysis could add to the existing knowledge needed for AFIT/CREATE management and CREATE management as a whole.

Research Objectives

The objectives of the proposed research were to:

1. Refine the Anderson and Purnell model, using an extended data base, to allow forecasting of computer usage.

2. Forecast the computer usage for AFIT/LS for a period of twelve months.

3. Determine if the model could be used for specification purposes for other computing systems.

Research Questions

1. Can the Anderson and Purnell model be further developed for use as an accurate forecasting tool for AFIT/CREATE support requirements?

a. If not, what factors contributed to the model not providing accurate forecasting?

b. If so, what level of use will the model project for a period of twelve months?

2. Can the model be used to define future system requirements for AFIT?

3. Does the model have a wider application for use in future computer system specifications?

CHAPTER II

METHODOLOGY

Collection of Data

General

The specific computer operating data available from the accounting report are[1]:

VARIABLE		MNEMONIC
Number of Log-Ons		LOGONS
Number of Batch Jobs		BAJOBS
CPU Hours Batch	Prime	CPUHRBP
CPU Hours Batch	Non-Prime	CPUHRBNP
CPU Hours Time Sharing	Prime	CPUHRTP
CPU Hours Time Sharing	Non-Prime	CPUHRTNP
Core Hours	Prime	COHRSP
Core Hours	Non-Prime	CPUHRSNP
Tape IO Hours	Prime	TAPEHRP
Tape IO Hours	Non-Prime	TAPEHRNP
Log-On Hours	Prime	LOGHRSP
Log-On Hours	Non-Prime	LOGHRSNP
Number of Lines CARDIN[2]		LNCARDIN

These data were collected by two methods:

1. Data for the period January 1974 to June 1976 already extracted from the monthly CREATE accounting reports for previous research were recovered from tape storage.

---

[1]A comprehensive list of all variables used in this research, and their mnemonics, is given in Appendix A.

[2]CARDIN is a time-sharing subsystem of CREATE that allows programs that are available on time-sharing files to be processed as batch jobs and their output sent to the batch-remote printer.

2. Data for the period July 1976 to March 1977 were extracted from monthly CREATE accounting reports. These data were used to extend the data base previously used.

For each parameter, the monthly CREATE accounting report provides usage data by categories of system users. The users of the AFIT/CREATE system are:

PROBLEM NUMBER	USER FACILITY	DESCRIPTION
WP1186	LSS	School of Systems and Logistics (Support)
WP1187	LSC	Faculty Support (Continuing Education)
WP1188	LSC	Student Classroom Support (Continuing Education)
WP1189	LSG	Faculty Support (Graduate Logistics)
WP1190	LSG	Student Thesis Support (Graduate Logistics)
WP1191	LSG	Student Classroom Support (Graduate Logistics)

Data Manipulation

For purposes model development, prime and non-prime data were aggregated. The distinction between prime and non-prime is made to meet the "billing" requirements of AFLC Regulation 400-25 (11). Such distinction between the two for the purposes of forecasting monthly requirements was not considered necessary.

Data were also aggregated to provide a single monthly total for each computer operating parameter without distinguishing between the different categories of users. Aggregation did not appear to influence the validity of data nor any relationships between variables.

The aggregated variables used in forecasting model were:

VARIABLE	MNEMONIC
Number of Log-Ons	LOGONS
Number of Batch Jobs	BAJOBS
CPU Hours Batch	CPUHRB
CPU Hours Time-Sharing	CPUHRT
Core Hours	COHRS
Tape IO Hours	TAPEHRS
Log-On Hours	LOGHRS
Number of Lines CARDIN	LNCARDIN

Model Validation

The first step in this research effort was to validate the conclusions of the Anderson and Purnell (1) research. Their research made use of Honeywell's Time Series Forecasting Program (TCAST) for analysis of past data and synthesis of the analysis to form a forecast. Validation was accomplished by comparing the forecasts of computer resource requirements to actual usage using TCAST.

It was concluded by Anderson and Purnell (1) that time-series forecasts could be used exclusively to predict AFIT computer usage. However, a change in the schedule of the Graduate Logistics classes in 1975 produced variations in the data base. This change contributed to an unacceptably large Mean Absolute Deviation (MAD)[3] (1:59-68). However, no

---

[3]MAD is defined as the average of the sum of the absolute differences between the actual observed value and the model fitted value.

comparison was made between actual forecasts and subsequent usage rates that were not included in the base series.

To validate the Anderson and Purnell conclusions, monthly time-series forecasts were made for each computer parameter. Then, these forecasts were compared to the actual observed value of the parameter.

Anderson and Purnell concluded that time-series forecasting could be used to forecast computer usage. To validate the forecasting technique employed, an analysis of the forecast error was undertaken. Validation was attempted for both short-term and long-term forecast lead-times.

A short lead-time was dictated by the shortest practical time that may elapse between making a final commitment and the time that the commitment is felt. For the AFIT/CREATE system, short lead-time was estimated to be two time periods (months). Data for January usage is not available until mid-February and therefore is not available for providing a short-term forecast for February. However, January's data may be added to the existing data base to forecast usage for March, giving a lead-time of two months.

Long-term forecasting allows management to plan the use of available resources. With the past data (30 months) a long-term forecast of 12 months is the maximum that could be expected with any accuracy because noise in the observed data is amplified by the forecast. The longer the lead-time, the greater the amplification (2:214).

Each forecast estimate was made using optimum forecast parameters that minimize the MAD for the lead-time specified. The necessary margin for the error of an estimate was established as a multiple (K) of the MAD. A confidence interval for each estimate was established using the formula:

Confidence Interval = $\bar{y} \pm K(MAD)$

where $\bar{y}$ is the mean of the observed data.
K is the safety factor for a given level of confidence (i.e. the probability that the predicted interval will not fail to contain the actual value).

If the predicted values occur within the confidence interval, then the technique was not necessarily considered to be a valid forecasting model of AFIT/CREATE requirements. The size of the MAD compared to the mean of the variable being forecast was also assessed. The smaller the ratio of the MAD to the mean, the better the forecasting model.

Further Model Development

General

The previous data base used by Anderson and Purnell included data for the period January 1974 to June 1976. The computer usage data available from July 1976 to March 1977 were added to the data base to provide an extended time dependent series.

The detection of bad data and subsequent correction was the first step. Where inconsistencies appeared to exist,

the activities for that period were examined. Judgements concerning the validity of data and the actions taken to correct deficiencies depended upon the situation encountered.

Model development initially prodeeded on the assumption that the change in schedule of the two Graduate Logistics classes did not affect any characteristic inherent in the data. Then data preceeding March 1975[4] (time of schedule change) was removed from the data base. The models for the two sets of data were compared.

If the model without the pre-March 1975 data provided better accuracy compared to the model developed using the pre-March 1975 data, it was to be concluded that the change in the schedule had a significant influence on the usage pattern of the AFIT/CREATE system. In this case, model development proceeded without the pre-March 1975 data. Because of the results obtained, the June 1976 to March 1977 data was also used in a separate analysis. The results obtained with this series was then compared to the other series.

Before any statistical analysis was done, the data was visually inspected. All aggregated computer usage parameters to be placed in the model were plotted on a

---

[4]Initially, it was intended to make May 1975 as the separation point because the new Graduate Logistics schedule began in May 1975. However, March 1975 was chosen to give 24 months of data; April 1975 to March 1977.

common time-series axis. The data were examined to determine if a relationship between two or more variables existed. For example, a relationship between LOGONS, LOGHRS and CPUHRT could be expected to exist. Did the plot indicate an increase in LOGONS is accompanied by an increase or decrease in LOGHRS or CPUHRT or both?

In developing a forecasting model, it was considered necessary to have an understanding of the factors influencing forecasting accuracy. Brown (2) discussed three different stochastic elements of errors encountered in problems of forecasting that affect accuracy.

1. The basic element is <u>noise</u>, which obscures the true process underlying the sequence of observations. The noise in a set of observations will prevent obtaining exact values of the true coefficients for the process being forecast. This noise is defined as:

$$x(t) = \xi(t) + E(t)$$

where $x(t)$ is the observed value.
$\xi(t)$ is the true process.
$E(t)$ is the noise at time period t.

2. The second stochastic element is the <u>residual</u>; the difference between a model and the actual observation. The model fitted to the observations may not be an accurate representation of the true process. The coefficients in the true process may also be changing slowly with time so that the current process is not the same as the process generating earlier observations.

These factors combine to create the residual at time T-j defined by:

$$e(T-j) = x(T-j) - \underline{\hat{a}}'(T)\underline{f}(-j)$$

where T .... is the current time period.
j .... is the number of time periods back from T.
x(T-j) is the observed value at time T-j.
$\underline{\hat{a}}'(T)$ is the row vector of the estimated value of the coefficients for time T.
$\underline{f}(-j)$ is the vector of the fitting functions[5] at time -j.

3. The third stochastic element is <u>forecast</u> <u>error</u>. Forecast error is like a residual in that it is the difference between the actual observed values and the forecast values. The distinction is that the forecast is forward in time so that the future observation is not one of the observations used in estimating the model coefficients. The forecast error is defined as:

$$e(T+t) = x(T+t) - \underline{\hat{a}}'(T)\underline{f}(t)$$

where T .... is the current time period.
t .... is the lead-time for the forecast.
e(T+t) .. is the forecast error at time T+t.
x(T+t) .. is the observed value at time T+t.
$\underline{\hat{a}}'(T)$ ... is the row vector of the estimated value of the coefficients for time T.
$\underline{f}(t)$ .... is the vector of the fitting functions.
$\underline{\hat{a}}'(T)\underline{f}(t)$ is the vector product representing the model of the observations from time period 1 to T.

The forecasting process amplifies the noise. That is, there is a distribution of the forecast itself caused by

---

[5]The fitting functions used were the constant, linear and quadratic functions available using TCAST. Higher order polynomials were not available and were not be employed. The method TCAST uses to calculate the optimum fitting function is dealt with in detail commencing on page 25, Trend Analysis.

past noise. The distribution of the forecast errors is the convolution of the noise distribution and the forecast distribution (2:267-268).

Correlation

Bivariate correlation provides a single number (correlation coefficient) which summarizes the linear relationship between two variables. These correlation coefficients indicate the degree to which variation (or change) in one variable is related to variation (or change) in another variable. A correlation coefficient not only summarizes the strength of an association between a pair of variables, but also provides a means for comparing the strength of a relationship between one pair of variables and a different pair (7:276).

Pearson's bivariate correlation analysis was performed using the Statistical Package for the Social Sciences (SPSS) (7). Scattergrams were also produced to give presentations of the relationships between the selected pair of variables. The principle assumption of correlation analysis is that the variables have a joint bivariate normal distribution. Therefore, it was assumed that variables to be introduced into the model were random variables with a joint bivariate distribution. Pearson's product-moment correlation coefficients, symbolized by r, and scattergrams were produced for all variable pairs.

Pearson's r, takes on a value of +1.0 to -1.0. The larger the absolute value of r, the stronger the linear relationship between the two variables. If r is positive, the two variables tend to increase (or decrease) together. A negative r denotes an inverse relationship--as one variable increases the other tends to decrease.

To test the statistical significance of the population correlation coefficient (ρ_{xy}) the hypothesis to be used is:

$$H_0 : \rho_{xy} = 0$$

$$H_1 : \rho_{xy} \neq 0$$

SPSS reports the significance tests for each coefficient r (r being an estimate of the population parameter ρ_{xy}). The level of significance is derived from the use of the Student's t distribution with n-2 degrees of freedom for the computed quantity, and is calculated by:

$$t_s = r \left[\frac{n - 2}{1 - r} \right]^{1/2}$$

where n is the number of points correlated.

Regression

Following the examination of the data on a common time axis and the use of correlation analysis, linear regression techniques were applied. Both Simple Linear Regression (SLR) and Multiple Linear Regression (MLR)

techniques were applied using SPSS. SLR is used to determine the relationship between a dependent or criterion variable and one independent variable. MLR is used to determine the relationship between a dependent or criterion variable and a set of independent or predictor variables. The SPSS multiple linear regression subprogram REGRESSION (used for both SLR and MLR) combines standard multiple regression and stepwise procedures in a manner which provides control over the inclusion of independent variables in the regression equations (7:320-322). SPSS uses the method of least-squares for calculating the regression line.

A measure of efficiency of the regression line is referred to as the _coefficient of determination_ and is calculated by:

$$R^2 = \frac{EV}{TV}$$

where $TV = \sum(y-\bar{y})^2$; the total variation of the dependent variable about its mean; $\bar{y}$.

$EV = \sum(\hat{y}-\bar{y})^2$; the variation explained by the regression line.

If $R^2=1$, then all the variation is explained by the regression, and a "perfect" fit between the dependent and independent variable(s) has been shown.

For MLR, SPSS provides for the independent variables to be entered into the equation in a predetermined order or by _forward_ (_stepwise_) _inclusion_. The former procedure is used when there is a definite causal ordering among the variables. No causal ordering for the computer usage

parameters was assumed, and forward (stepwise) inclusion was used in MLR.

Using forward (stepwise) inclusion, the variable that explains the greatest amount of variance (unexplained by the variables already in the equation) enters the equation at each step. The independent variable chosen for entry is the one which has the largest squared partial correlation coefficient with the dependent variable, and is defined by:

$$r^2_{y1.23} = \frac{R^2_{y.123} - R^2_{y.23}}{1 - R^2_{y.23}}$$

where $r^2_{y1.23}$ is the square of the partial correlation of y and x_1 with x_2 and x_3 in the equation.

$R^2_{y.123}$ is the coefficient of determination with x_1, x_2 and x_3 in the equation.

$R^2_{y.23}$ is the coefficient of determination with x_2 and x_3 in the equation.

The SPSS subprogram enters the variables in single steps from best to worst provided the variable meets the established statistical criteria[6]. If the statistical criteria are not met, an independent variable will never be entered into the equation.

The F ratio is computed as a test for the significance of a regression coefficient. The F ratio for a

---

[6]The statistical criteria used were the F ratio and the tolerance, T. The default values for F and T used for these analyses were F=0.01 and T=0.001.

given variable is the value that would be obtained if that variable were brought in on the very next step. The F ratio is calculated by:

$$F^{1}_{n-p} = \frac{b_j^2}{s_{b_j}^2}$$

where n is the number of data points.

p is the number of terms (variables plus constant) in the regression equation.

b_j is the estimate of the coefficient of the independent variable.

s_{b_j} is the standard deviation of the coefficient b_j.

The test on the statistical significance of the presence of an independent variable was conducted in isolation--without testing any other independent variable in that step.

A second condition to be met before an independent variable is entered in to the regression equation is the tolerance (T). The tolerance of an independent variable being considered for inclusion is the proportion of the variance of that variable not explained by the independent variables already in the equation (7:346). If the tolerance criteria[7] are not met, the independent variable will never enter the equation.

---

[7]T has a possible range from 0 to 1. A tolerance of 0 would indicate that a given variable is a perfect linear combination of the other independent variables. A tolerance of 1 would indicate that the variable is uncorrelated with the other independent variables. An intermediate value of 0.6 would indicate that 60% of the variance of a potential independent variable is unexplained by the variables already entered.

If an independent variable lacked statistical significance, it was analyzed based on the following criteria:

1. The data base was too small to give insight into true relationships among variables.

2. Multicollinearity[8] was present. This was determined using the tolerance value. If the tolerance value suggested multicolinearity, two alternatives were evaluated:

a. If it was believed that the variable should be included, a new variable was created as a composite of the intercorrelated variables. The new scale variable was introduced into the equation.

b. If the introduction of the variable into the equation was not considered necessary, only one of the variables in the highly intercorrelated set was used to represent the common underlying dimension (7:341).

3. The independent variable did not contribute. If the model was acceptable without the independent variable, it was excluded from the regression.

In addition to the significance tests described above, the overall model was evaluated by:

---

[8]Multicollinearity is the inter-correlation between two or more independent variables in a MLR model. It reduces the ability to account for the explanatory power of the particular independent variable in the model.

1. Evaluating the statistical significance of the coefficient of determination using the hypothesis:

$$H_0 : R^2 = 0$$

$$H_1 : R^2 \neq 0$$

The test statistic uses the F test and was calculated by:

$$F^{p-1}_{n-p} = \frac{(EV)(n-p)}{(UV)(p-1)}$$

2. Evaluating the <u>standard error of the estimate</u> ($s_{y.x_1x_2...x_n}$) as a measure of the distribution of the dependent variable given the values for the independent variables. The standard error of the estimate is defined by:

$$s_{y.x_1x_2...x_n} = \left[\frac{(UV)}{(n-p)}\right]^{1/2}$$

The lower the standard error of the estimate, the better the regression line represented the true relationship between the variables.

3. Subjectively assessing the appropriateness of the regression as a means of forecasting computer resource requirements.

4. In the case of SLR, a confidence interval which was calculated for both the regression line and the dependent variable.

<u>Time-Series Forecasting</u>

Time-series forecasting techniques may be applied when a process is observed at discrete sampling intervals. Since the data base was collected at discrete monthly intervals, it was considered appropriate to apply

time-series forecasting of the computer operating parameters. In addition, the annual rotation of Graduate Logistics courses and the computing workload within each course may exhibit cyclic effects of computer usage that are suitable for time-series forecasting analysis.

Honeywell's TCAST program was used for time-series forecasting analysis. TCAST follows four fundamental steps to provide predictions (3:1-1):

1. Cyclic analysis of past data.
2. Trend analysis of past data.
3. Error analysis for comparing forecast with actual data.
4. Synthesis of analysis to form a forecast.

<u>Cyclic Analysis</u>. TCAST provides the cycle length which yields the minimum relative error between the observed and forecast data. It was expected that the predominant cycle for AFIT/CREATE computer usage would be 12 months because of yearly course schedules including a heavier demand for computer support as students move toward completion of their research thesis. However, this was not the case. No variable gave a cycle of 12 or a factor thereof. In some cases, the cycle selected by TCAST was used in the analyses. If the cycle selected by TCAST was one, the cycle that gave the next smallest cyclic error was used.

<u>Trend Analysis</u>. TCAST uses exponential smoothing for forecasting prevailing general tendencies (trends), and for

determining relationships between past and future data. The orders of smoothing used are:

1. First Order. First order exponential smoothing provides the weighted moving average at a particular time as the forecast for all future time periods. The basic equation for first order exponential smoothing is:

$$S1_t = \alpha x_t + (1 - \alpha) S1_{t-1}$$

where $S1_t$ is the first order exponentially smoothed average of the observations through time t.
x_t is the observed value at time t.
α is the smoothing constant; $0 < \alpha < 1$.
$S1_{t-1}$ is as above through time t - 1.

The data are weighted to give more or less weight to older data by the alpha value. The data p periods ago are weighted by:

$$\alpha (1 - \alpha)^p$$

2. Second Order. Second order exponential smoothing takes into account the linear rate of change in the single exponentially smoothed average. The basic equation is:

$$S2_t = \alpha S1_t + (1 - \alpha) S2_{t-1}$$

where $S2_t$ is the second order exponentially smoothed average of the observations through time t.
$S2_{t-1}$ is as above through time t - 1.

3. Third Order. Third order smoothing takes into account the quadratic rate of change in the exponentially smoothed average. The basic equation is:

$$S3_t = \alpha S2_t + (1 - \alpha) S3_{t-1}$$

where $S3_t$ is the third order exponentially smoothed average of the observations through time t.
$S3_{t-1}$ is as above through time t - 1.

TCAST was used to determine the optimum alpha and order of smoothing to minimize the MAD. The optimum alpha and order of smoothing vary for a specified lead-time. The lead-times were:

1. Short-Term. A lead-time of two periods (months) was used for short-term forecasting.

2. Long-Term. A maximum lead-time of 12 months was established for long-term forecasting.

Error Analysis. The alpha and type of smoothing which yield the most accurate forecast (minimum MAD) was used for model development. This error is minimized in the forecast over the specified lead-time.

Synthesis. After all analyses have been made, the results from each are synthesized by TCAST to form a composite forecast (3:3.1-3.6).

Model Validation

Once the statistical tests and analyses were completed and a model developed for AFIT/CREATE, the same techniques were applied to an additional data base; CREATE as a total system. The same statistical techniques and decision processes used for AFIT/CREATE were applied to CREATE.

Evaluation Criteria

Either regression or time-series forecasting, or a combination of the two techniques was used to forecast

computer usage requirements for the AFIT/CREATE system. In addition to the statistical analysis described above, the following criteria were evaluated:

1. Accuracy: The more accurate model was preferred.

2. Simplicity: A technique that management is able to comprehend would be preferred even if some accuracy must be sacrificed.

3. Appropriateness: The model should be able to meet the needs of the real world environment.

4. Cost: The benefits of introducing and employing the model in the real world should be at least commensurate with the cost of introducing the model and maintaining the data base.

List of Assumptions

Collection of Data

1. The distinction between prime and non-prime accounting data was not considered necessary to forecast monthly requirements of computer usage.

2. The aggregation of data for the different categories of users within AFIT/LS did not influence the validity of the data or any relationship between variables.

3. No errors exist in the data base for the period Jan 74 to Jul 76 compiled by Anderson and Purnell.

Further Model Development

1. The change in schedule of the two Graduate Logistics classes does not affect the usage pattern of the AFIT/CREATE system.

2. For the purpose of correlation analysis, the variables are random variables with a joint bivariate normal distribution.

CHAPTER III

DATA COLLECTION AND MANIPULATION

Graphic Presentation

Presentation of the data in graphic form is given in Figures 2A through 2D for the period of January 1974 to March 1977. The variables given in each figure are:

Figure	Variables[1]
2A	BAJOBS, CPUHRB and LOGONS
2B	CPUHRB, COHRS and TAPEHRS
2C	LOGHRS, LOGONS and CPUHRT
2D	COHRS, TAPEHRS and LNCARDIN

The graphs give a visual indication of the relationship between the computer operating parameters being studied as well as the behavior of each parameter versus time. In Figure 2C, the close positive correlation between LOGHRS and LOGONS can be seen. Examination of the other graphs indicate similar relationships.

The Anderson and Purnell (1) research did not include LNCARDIN. However, since these data were readily available from the monthly accounting records, and the CARDIN system is used extensively by faculty and students, it was considered necessary to include LNCARDIN information.

---

[1]The variables have been scaled so that they may be presented on a common axis.

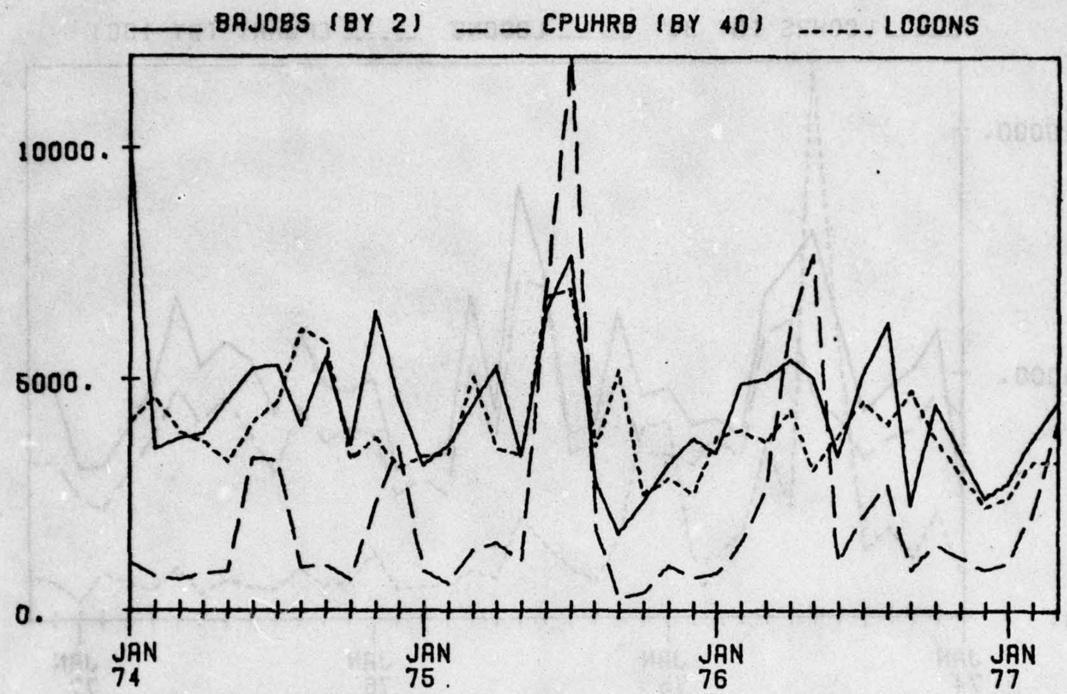


Figure 2A. GRAPH: BAJOBS, CPUHRB AND LOGONS--AFIT/CREATE

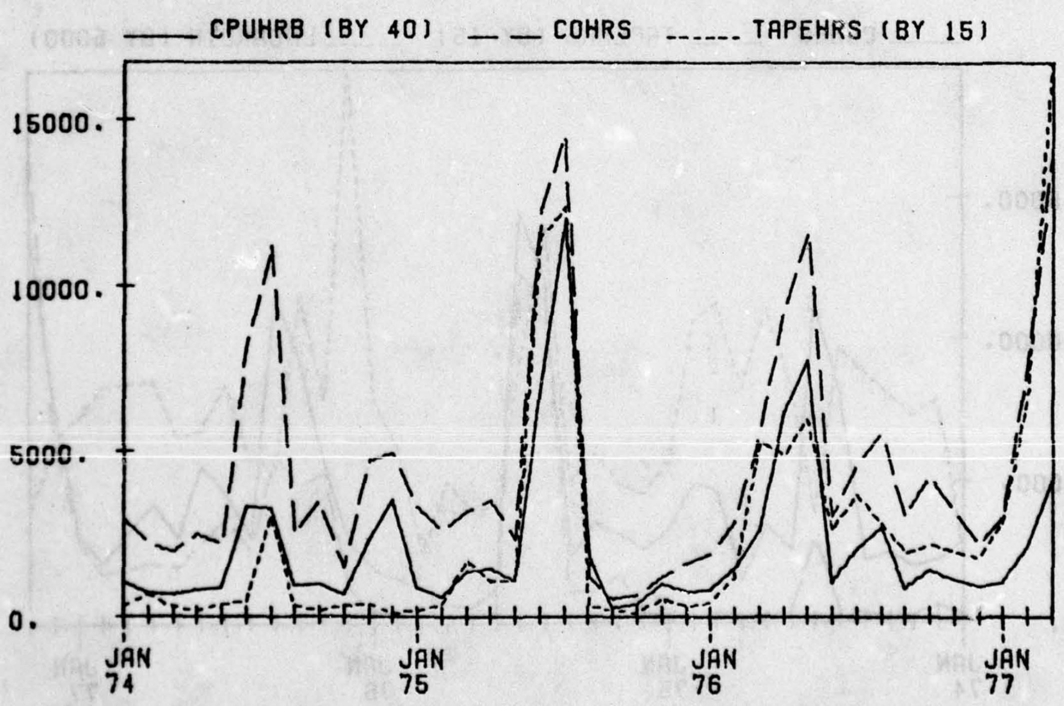


Figure 2B. GRAPH: CPUHRB, COHRS AND TAPEHRS--AFIT/CREATE

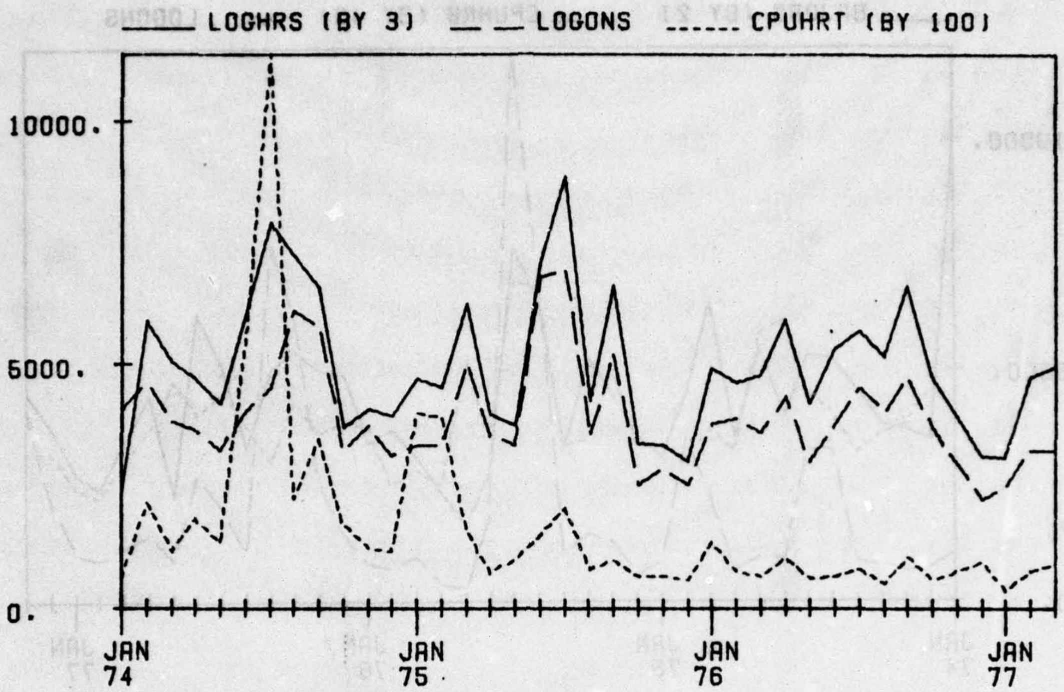


Figure 2C. GRAPH: LOGHRS, LOGONS AND CPUHRT--AFIT/CREATE

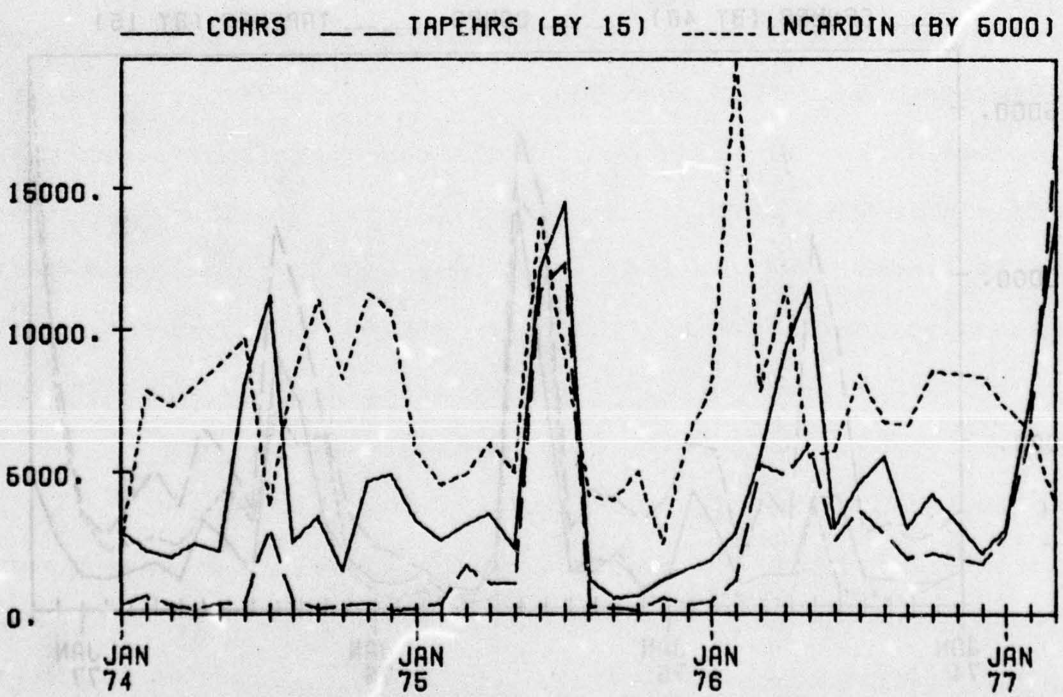


Figure 2D. GRAPH: COHRS, TAPEHRS AND LNCARDIN--AFIT/CREATE

The data are given in Appendix B, and contain the monthly aggregation of prime and non-prime variables for the six problem numbers given in Chapter II.

Interpretation and Correction of Data

It was originally intended to validate the Anderson and Purnell (1) research without changing the data they used. However, the graphical presentation of the data posed questions concerning data accuracy. These questions were resolved before validation was attempted.

Figure 2C shows a high positive correlation between LOGHRS and LOGONS. However, the December 1974 value for LOGHRS indicated a prominant increase, while LOGONS actually decreased. A search of the accounting records for December 1974 indicated that LOGHRS were in error and the data base was corrected.

The accounting records available for November 1976 only contained data for the total CREATE usage. No data were available for individual problem numbers. To provide data for this period, the data available for October 1976 and December 1976 were averaged.

When the data for December 1976 were introduced, all computer variables indicated a significant increase for that months usage. This was contrary to the behavior expected since students were doing little computing work during the month and were on leave for two weeks. Also, it was noticed

that the LOGONS, while expected to be whole numbers, were actually multiples of 10 for all problem numbers. Therefore, all data for December 1976 were divided by a factor of 10. This made the December 1976 figures more consistent with the other data points as shown in Figure 2. Also, AFLC CREATE Management verified that the December 1976 figures had been inadvertently multiplied by 10.

No data could be found for July 1975. Yet, the Anderson and Purnell research included data for that period. When introducing LNCARDIN data, the values for June 1975 and August 1975 were averaged to obtain a data point for July 1975.

It was observed that the values for BAJOBS and LOGONS for July 1975 were exactly the same as for June 1975. However, the value for other variables for that same time period were different. Since BAJOBS and LOGONS are contained in a separate part of the accounting data, it was assumed that Anderson and Purnell had found the July 1975 data with the exception of BAJOBS and LOGONS. All of the data, with the exception of BAJOBS and LOGONS, showed distinct peaks for July 1975.

These two variables were adjusted as follows:

(1) BAJOBS. COHRS gave the highest bivariate correlation with BAJOBS. Therefore, the ratio between BAJOBS and COHRS for June 1975 was multiplied by the value of COHRS for July 1975 to give a value for BAJOBS for July 1975.

(2) LOGONS. LOGHRS gave the highest bivariate correlation with LOGONS. Therefore, the ratio between the means for LOGONS and LOGHRS was multiplied by the value of LOGHRS for July 1975 to give a value for LOGONS for July 1975.

No data were available for April 1977 or May 1977 because of failure of the CREATE accounting system. The data base for this research was therefore begun at January 1974 and terminated at March 1977.

CHAPTER IV

RESULTS

Validation of Anderson and Purnell Research

Validation of the Anderson and Purnell (1) research was done using TCAST with alpha values to an accuracy of two decimal places, except where the MAD error was not significantly different between two alpha values. Then a higher accuracy for alpha was used. In addition, the value for alpha less than Ø.6667[1] which gave the minimum MAD was selected. A lead time of two months was used.

For each variable, a forecast was made for July 1976, using the Anderson and Purnell data base for the period January 1974 to May 1976. An additional data value was successively added to the data base, and another forecast was made. Forecasts were made up to March 1977 and then compared with actual observations[2].

The cyclic analysis of TCAST provides for automatic selection of the cycle length which provides minimum cyclic

---

[1]Values of alpha less than 2/(L+1), where L is the lead time, must be used. Otherwise, the basis for the forecasts becomes too heavily biased by the most recent data (3:3-6).

[2]The complete results of this validation process are given in Appendix C.

error--the relative measure of residual variance for that cycle length. While the data in Figure 2 indicate dominate cycles by visual inspection for most variables, TCAST was unable to identify the cycle and gave a dominant cycle of one in a number of the forecasts.

Figure 3A indicates the forecast that would occur using the TCAST generated cycle of one. Instead of using the cycle of one, the cycle that gave the next minimum relative error was used. In some cases, the difference in relative error between the two cyclic components was as low as 0.01%, and for others, as high as 10%.

However, to introduce some cyclic component was considered to be more appropriate than to introduce no cycle at all. For the variable COHRS, shown in Figure 3A, the same forecast was accomplished using a cycle of 10. This is displayed in Figure 3B. A comparison of the two forecasts for January 1977 are[3]:

	CYCLE=1	CYCLE=10
CYCLIC ERROR	463961	476241
ALPHA	0.01	0.009
TYPE SMOOTHING	1	3
MAD	3283	2140
FORECAST	3323	8427
ACTUAL	2999	2999

[3]The methodology followed by Anderson and Purnell in their research to overcome this type of problem is not known.

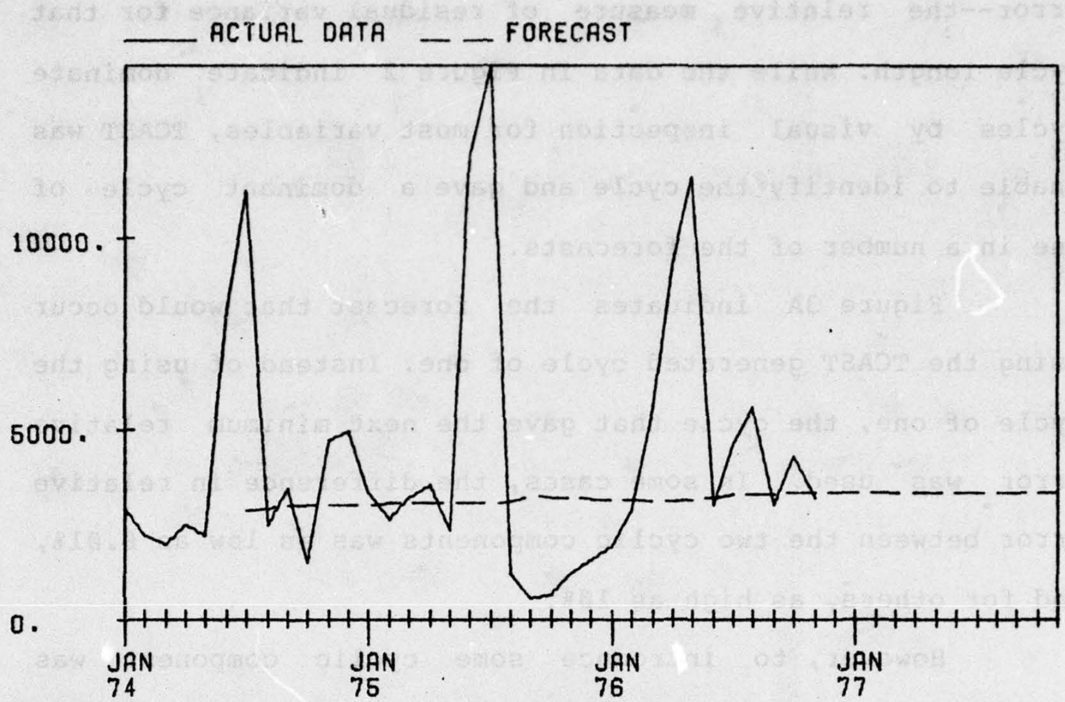


Figure 3A. TIME-SERIES FORECAST--COHRS--USING TCAST CYCLE

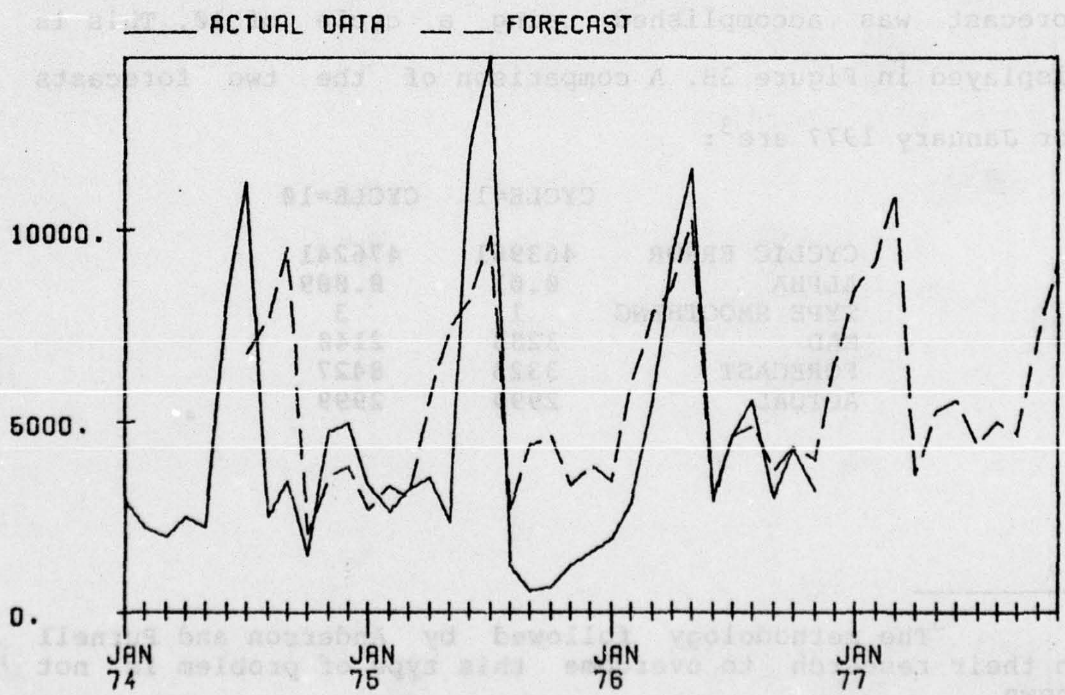


Figure 3B. TIME-SERIES FORECAST--COHRS--USING CYCLE OF 10

Review of the results in Appendix C reveals several significant factors:

1. An overwhelming majority of the forecasts were more than 25% larger or smaller than the actual value.

2. A very large MAD was experienced in most cases. For 90% confidence, the half width confidence interval is 2.062 times the MAD (2:286). With the large MAD, a realistic confidence interval could not be established.

3. A number of variables did give a consistent, or near consistent cyclic component for all forecasts. These were:

Variable	Cycle
LOGHRS	12
CPUHRB	10
BAJOBS	5

4. The more data points added to the forecast, the better TCAST was able to identify a cyclic component. The analysis of LOGONS, TAPEHRS and COHRS indicates this result. However, this was not attributed totally to the longer data base. It was due in part to more consistency in the cyclic components of the more recent data. Figure 3B indicates this characteristic. The composite peaks coincide with the actual data peaks at the end, but not at the beginning of the data.

5. CPUHRT was an exception to both of the above observations. It was characterized by a decline in usage as shown in Figure 2C.

Further Model Development

General

It was intended to compare the results of model development by doing separate analyses with and without pre-May 1975 data. However, since data was not available for April and May 1977, the initial comparison was done using April 1975 as the cut-off. This allowed the use of a full 24 months of data for the period April 1975 to March 1977 to compare with the complete data base for the period January 1974 to March 1977.

However, several early observations concluded that separate model development using the two different time-series was not necessary. The results of those observations were:

1. During validation of the Anderson and Purnell research, it was noted that the cyclic patterns were more evident in the latter data. Peaks occurred in July 1975 and July 1976. The July 1974 data were not as consistent.

2. Pearson's bivariate correlation coefficients were calculated for both the January 1974 to March 1977 and April 1975 to March 1977 data. All coefficients were higher and levels of significance smaller with the pre-April 1975 data removed.

3. Autocorrelation coefficients were calculated for both time series. In general, removal of the pre-April 1975 data provided higher autocorrelation coefficients.

Data Distribution

Figures 2A to 2D give a visual indication of the variability of the parameters studied in this research. The mean and standard deviation for all parameters were calculated to provide a quantitative measure of this variability. For this research, three separate series were used. The mean and standard deviation for each series are as follows:

TABLE 1. COMPARISON OF MEANS AND STANDARD DEVIATIONS FOR AFIT/CREATE TIME-SERIES

Variable	Jan 74 - Mar 77		Apr 75 - Mar 77		Jun 76 - Mar 77	
	Mean	STDDEV	Mean	STDDEV	Mean	STDDEV
BAJOBS	2162.4	772.3	2052.6	747.8	1917.6	627.7
CPUHRB	53.7	59.5	64.5	71.4	43.7	25.8
CPUHRT	16.5	19.4	8.7	3.6	7.5	2.2
COHRS	4527.9	3713.1	4969.1	4248.7	5027.0	3478.8
TAPEHRS	166.0	247.0	245.1	286.8	286.8	305.0
LOGHRS	1652.3	480.5	1613.4	478.4	1580.7	369.1
LOGONS	3910.1	1059.2	3787.4	1202.1	3495.8	822.2
LNCARDIN	1496.6	650.8	1494.2	720.5	1413.7	297.4

No specific conclusions could be drawn from these results except that BAJOBS, CPUHRT, LOGHRS and LOGONS appeared to be decreasing with time, while COHRS and TAPEHRS were increasing. These observations were confirmed with later research. Additionally, high variability did exist for CPUHRB, COHRS and TAPEHRS.

Time Series Forecasting

The additional time-series forecasting analysis was done using a 24 month time-series; April 1975 to March 1977. Autocorrelation coefficients were calculated for each variable using a maximum lag value of 10 months. Shown as a continuous curve (although the function is discrete valued), the autocorrelation function for two variables are plotted in Figures 4A (LOGHRS) and 4B (CPUHRT). Figure 4A is interpreted as follows: The time-series values tend to be negatively correlated at a lag value of 5 months, positively correlated at a lag value of 7 months, and show no correlation at lags greater than 10 months.

TCAST was then used to establish a forecast for each variable using lead-times of 2 and 12 months. The forecasts consistently gave smaller MAD's than the January 1974 to June 1975 data-series used by Anderson and Purnell (TAPEHRS was an exception). A comparison of the MAD results are[4]:

	Apr 75 to Mar 77 MAD for a Lead Time of		Jan 74 to Jun 75 MAD for a Lead Time of 12
Variable	2	12	
BAJOBS	421	482	622
CPUHRB	25.9	24.8	33.6
CPUHRT	1.9	1.8	24
COHRS	1059	1618	3102
TAPEHRS	187	116	107
LOGHRS	238	304	350
LOGONS	530	769	909
LNCARDIN	408	544	-

---

[4] A comprehensive listing of the results obtained is given in Appendix D.

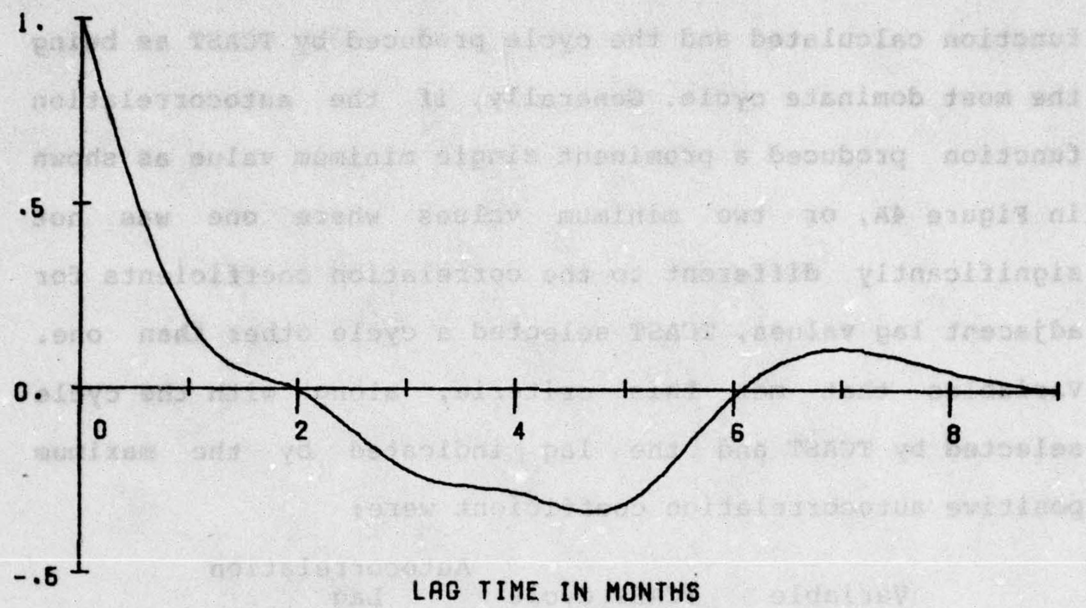


Figure 4A. AUTOCORRELATION FUNCTION--LOGHRS

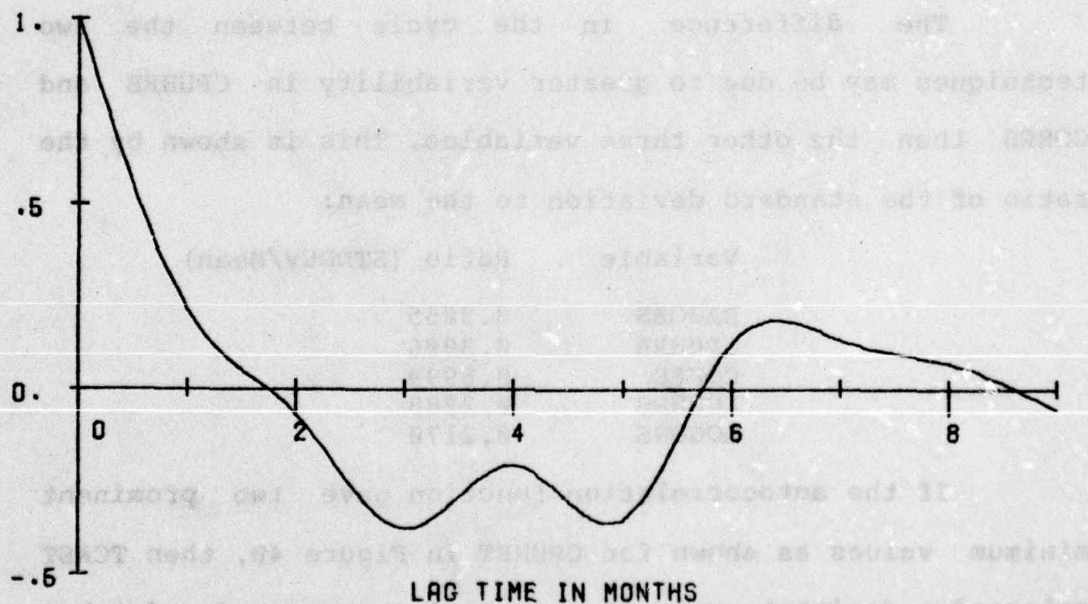


Figure 4B. AUTOCORRELATION FUNCTION--CPUHRT

A comparison was made between the autocorrelation function calculated and the cycle produced by TCAST as being the most dominate cycle. Generally, if the autocorrelation function produced a prominent single minimum value as shown in Figure 4A, or two minimum values where one was not significantly different to the correlation coefficients for adjacent lag values, TCAST selected a cycle other than one. Variables that met this criteria, along with the cycle selected by TCAST and the lag indicated by the maximum positive autocorrelation coefficient were:

Variable	TCAST Cycle	Autocorrelation Lag
BAJOBS	8	8
CPUHRB	10	8
COHRS	10	8
LOGHRS	7	7
LOGONS	7	7

The difference in the cycle between the two techniques may be due to greater variability in CPUHRB and COHRS than the other three variables. This is shown by the ratio of the standard deviation to the mean:

Variable	Ratio (STDDEV/Mean)
BAJOBS	0.3055
CPUHRB	0.3906
COHRS	0.6999
LOGHRS	0.2288
LOGONS	0.2170

If the autocorrelation function gave two prominent minimum values as shown for CPUHRT in Figure 4B, then TCAST selected a dominant cycle of one. Again, as was found when using the longer data-series to validate the Anderson and

Purnell research, the cyclic error for different lags was not significantly different. The results of using different cycles for CPUHRT (Mean=8.734) were:

Cycle	Cyclic Error	Smoothing Type	Alpha	MAD
1	0.9276	1	0.001	1.77
2	0.9753	1	0.001	8.01
7	1.0091	1	0.001	2.39

Hence, a simple weighted moving average was more appropriate than attempting to introduce a cyclic component. In fact, by introducing a cycle, the MAD could be increased significantly. Where TCAST gave a cycle of one, runs for a cycle of one and other cyclic components were made. The one that gave the minimum MAD was selected.

With one exception (TAPEHRS)[5], the MADs for the April 1975 to March 1977 time-series were smaller than those obtained with the pre-April 1975 data included. For the April 1975 to March 1977 time-series with a lead-time of 12 months, the confidence intervals were:

Variable	Mean	MAD	90% Confidence Limits Lower	Upper
BAJOBS	2052.6	482.4	1058.7	3046.5
CPUHRB	64.5	24.8	14.0	115.0
CPUHRT	8.7	1.8	5.0	12.4
COHRS	4969.1	1618.2	1632.8	8305.4
TAPEHRS	245.1	116.3	5.9	484.3
LOGHRS	1613.4	304.6	986.5	2240.2
LOGONS	3787.4	769.7	2201.7	5373.1
LNCARDIN	1537.7	544.1	415.8	2659.6

[5]TAPEHRS (Figure 2B) shows a significant increase for the last data point. This had the effect of introducing large residuals for the composite forecast.

While the confidence interval establishes the range within which a particular computing requirement can be expected, for management purposes the range is undesirably large in some cases. Figure 5 shows the confidence interval for LOGHRS (the ratio of MAD to mean is smallest for LOGHRS).

Correlation

Pearson's bivariate correlation analysis was done for all variable pairs in the following time-series[6]:

Jan 1974 to Mar 1977
Apr 1975 to Mar 1977
Jun 1976 to Mar 1977

Since the analysis was done on monthly data, aggregated for prime and non-prime values, a coefficient of at least 0.8 at a level of significance of 0.05 or less was assumed to indicate a "strong" linear relationship existed between the two variables correlated.

The pairs of variables that met this criteria for the April 1975 to March 1977 data were:

Variables		Coefficient	Significance
LOGHRS	LOGONS	0.949	0.001
COHRS	TAPEHRS	0.931	0.001
CPUHRB	COHRS	0.890	0.001

[6] A comprehensive listing of Pearson's bivariate correlation coefficients are given in Appendix E.

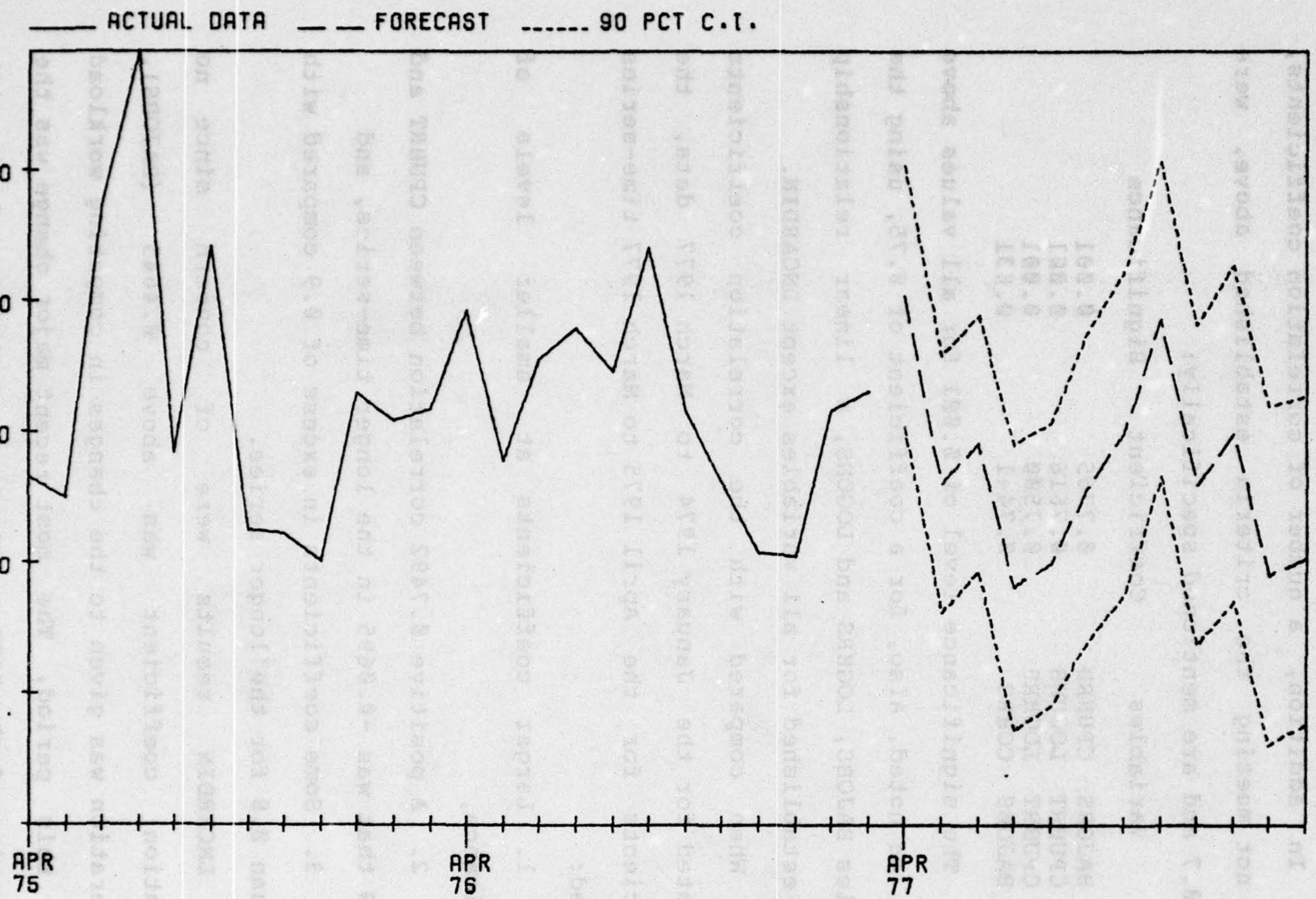


Figure 5. FORECAST AND CONFIDENCE INTERVAL--LOGHRS

In addition, a number of correlation coefficients, though not meeting the criteria established above, were above 0.7 and are mentioned specifically:

Variables		Coefficient	Significance
BAJOBS	CPUHRB	0.7765	0.001
CPUHRT	LOGONS	0.7616	0.001
CPUHRT	LOGHRS	0.7500	0.001
BAJOBS	COHRS	0.7441	0.001

The significance level of 0.001 for all values above should be noted. Also, for a coefficient of 0.75, using the variables BAJOBS, LOGHRS and LOGONS, a linear relationship can be established for all variables except LNCARDIN.

When compared with the correlation coefficients calculated for the January 1974 to March 1977 data, the coefficients for the April 1975 to March 1977 time-series provided:

1. Larger coefficients at smaller levels of significance,

2. A positive 0.7492 correlation between CPUHRT and TAPEHRS that was -0.0695 in the longer time-series, and

3. Some coefficients in excess of 0.9 compared with less than 0.9 for the longer series.

LNCARDIN results were of concern since no correlation coefficient was above 0.4645 (BAJOBS). Consideration was given to the changes in computing workload during this period. The most recent major change was the introduction of the SPSS statistical computing package with the commencement of the A Class of 1977 in May 1976. To

determine what relationship existed between the variables since the introduction of SPSS, the correlation results for the last 10 months data--June 1976 to March 1977--were compared to the results obtained for the two longer series.

The results were generally superior to those achieved using the two longer time-series. Some characteristics were:

1. Four coefficients were greater than 0.9.

2. LNCARDIN was correlated above 0.7 with two other variables; COHRS and TAPEHRS.

3. A number of smaller coefficients (less than 0.4) were evident whereas a large number of mid-range (0.4 to 0.7) coefficients existed in the longer time-series.

4. However, both BAJOBS and *CPUHRT did not have any* coefficients above 0.7 for the 10 month time-series.

In some respects, the longer time-series had advantages, while in others the shorter (10 month) time-series gave better results. These results will be discussed further in the regression research.

For the shorter time-series, the correlation coefficients above 0.7 were:

Variables		Coefficient	Significance
COHRS	TAPEHRS	0.9678	0.001
LOGHRS	LOGONS	0.9568	0.001
CPUHRB	COHRS	0.9372	0.001
CPUHRB	TAPEHRS	0.8642	0.001
COHRS	LNCARDIN	-0.7785	0.008
TAPEHRS	LNCARDIN	-0.7785	0.004

gave a good indication of relationships between the variables and reinforced our observations made by examining the variables on a common time axis. Figures 6 and 7 provide the results of two scattergrams for the 10 month time-series.

Figure 6 indicates the relationship between LOGHRS and LOGONS. A disappointing result was that obtained for BAJOBS and CPUHRB, which is shown in Figure 7. It had been thought that a stronger linear relationship between BAJOBS and CPUHRB would have existed.

Pearson's bivariate correlation was also done between the values for April 1975 to March 1976 and April 1976 and March 1977 for each variable. This was done to determine the correlation between the two consecutive groups of 12 months data. The results were:

Variable	Correlation Coefficient
BAJOBS	0.2125
CPUHRB	-0.0924
CPUHRT	-0.2809
COHRS	0.0059
TAPEHRS	0.1645
LOGHRS	0.5652
LOGONS	0.5667
LNCARDIN	-0.2609

Since the Graduate Logistics classes are scheduled at 12 month intervals, it was felt there could be a correlation between the two groups of data. However, the results indicate that no correlation exists. No conclusions could be drawn.

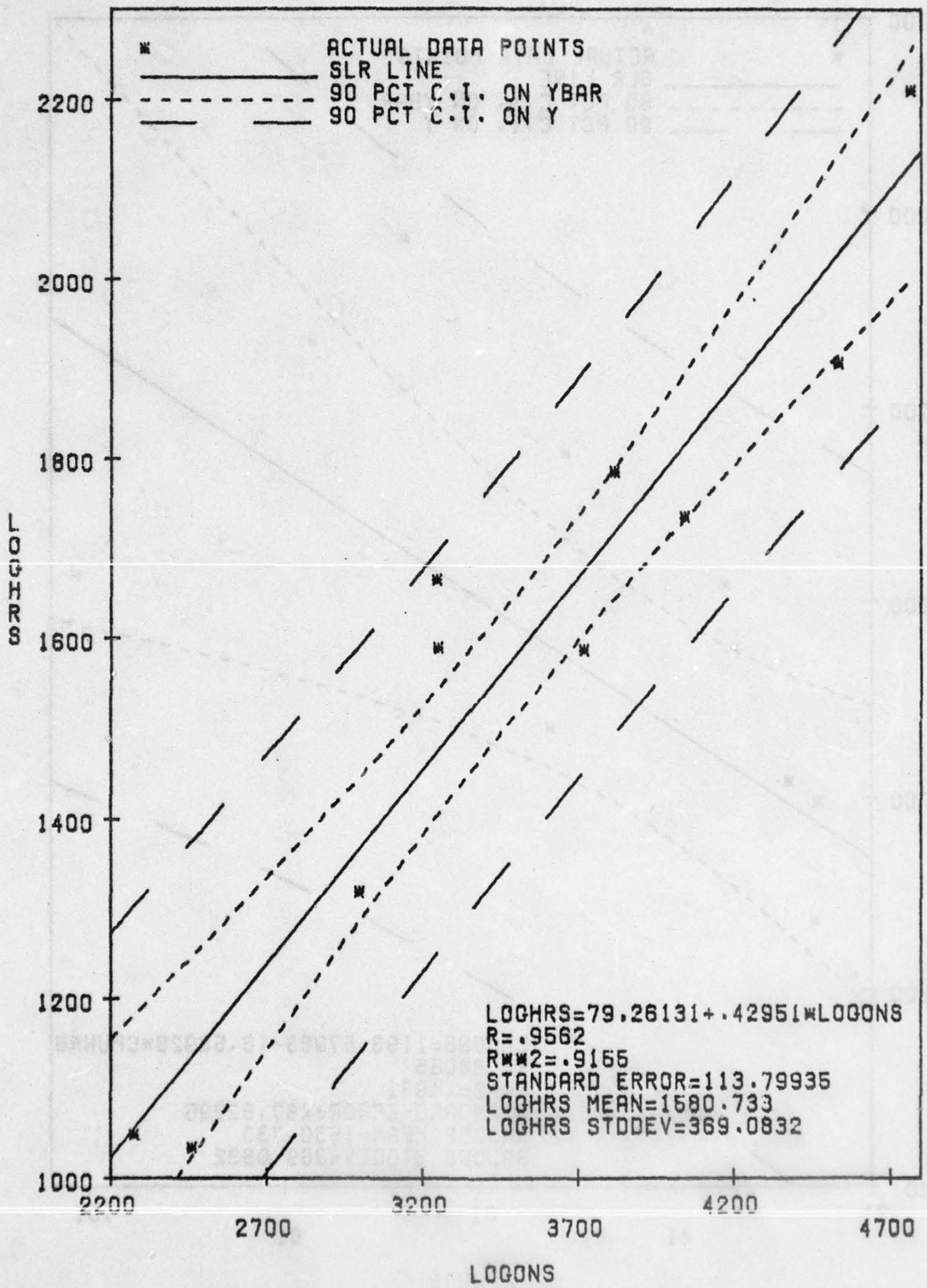


Figure 6. SCATTERGRAM, REGRESSION LINE AND CONFIDENCE INTERVAL--LOGHRS VERSUS LOGONS

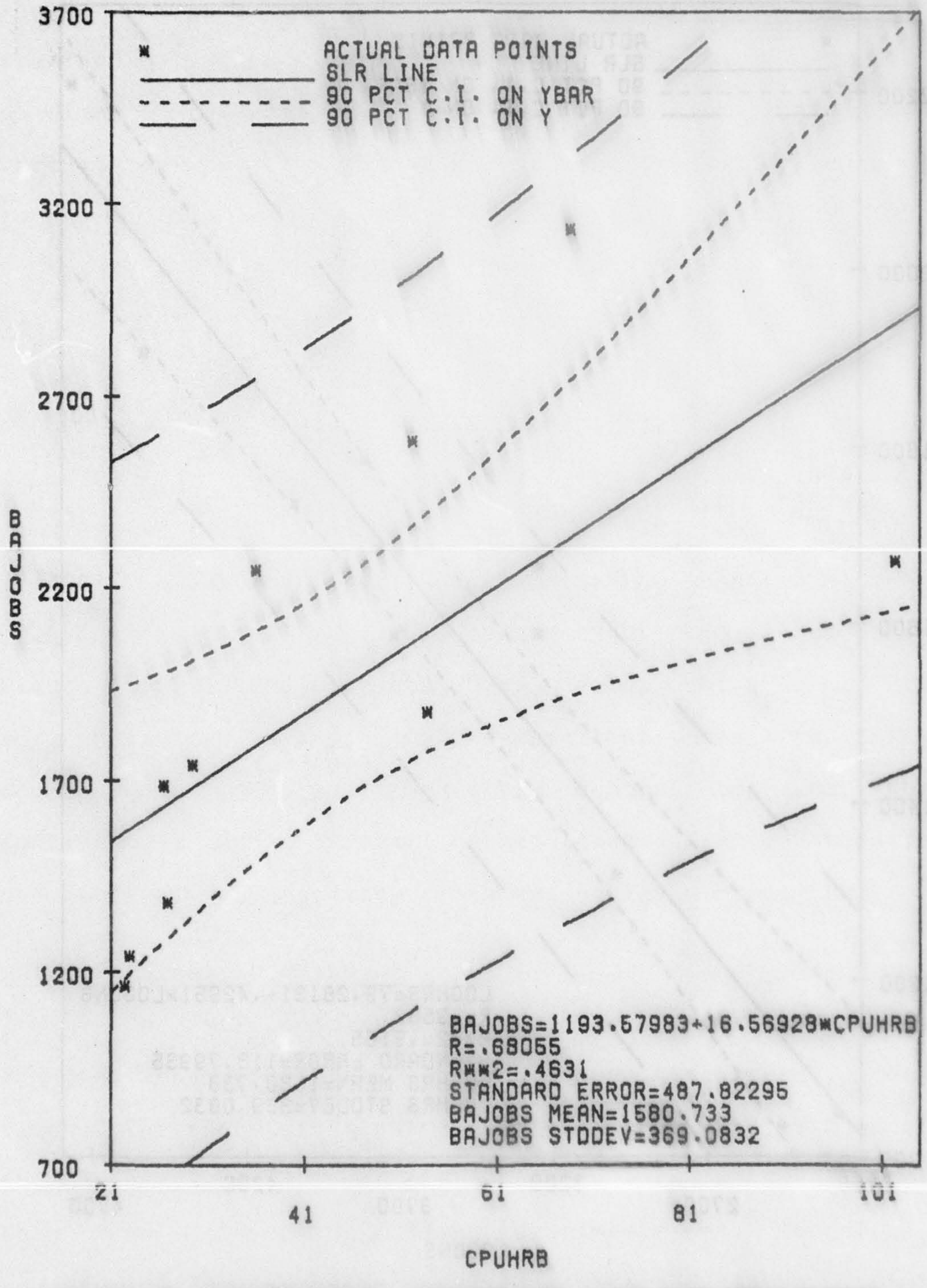


Figure 7. SCATTERGRAM, REGRESSION LINE AND CONFIDENCE INTERVAL--BAJOBS VERSUS CPUHRB

Simple Linear Regression

Initially, it was not intended to do simple linear regression (SLR). However, since the Pearson's bivariate correlation analysis produced better results than anticipated, SLR was attempted. A comparison of the results[7] for the April 1975 to March 1977 and June 1976 to March 1977 data series follows:

TABLE 2. COMPARISON OF SIMPLE LINEAR REGRESSION RESULTS FOR AFIT/CREATE TIME-SERIES

Dependent Variable	Jun 76 - Mar 77		Apr 75 - Mar 77		
	R^2	$s_{y.x}$	R^2	$s_{y.x}$	s_y
BAJOBS	0.585	481.8	0.461	487.8	627.7
CPUHRB	0.793	33.2	0.878	9.5	25.8
CPUHRT	0.580	2.4	0.132	2.1	2.2
COHRS	0.866	1586.0	0.937	929.5	3478.8
TAPEHRS	0.867	107.0	0.937	81.7	306.0
LOGHRS	0.901	154.2	0.915	113.8	369.1
LOGONS	0.901	387.6	0.915	253.5	822.2
LNCARDIN	0.216	652.4	0.606	198.0	297.4

Overall, both time-series produced approximately the same results for the coefficient of determination (R^2). However, the shorter time-series consistently gave a smaller standard error of the estimate ($s_{y.x}$), and therefore a smaller confidence interval for a given level of confidence. In most cases, the standard error of the estimate of the

---

[7]A comprehensive listing of the results of the SLR analysis for the June 1976 to March 1977 time-series is given in Appendix F.

shorter time-series was considerably smaller than the standard deviation of the dependent variable (s_y). The 90% confidence interval at the mean ($\overline{x}$) of the independent variable (x) for each dependent variable (y) was calculated to be:

Dependent Variable (y)	Independent variable(x)	Mean ($\overline{y}$)	Confidence Limits at $\overline{x}$ Lower	Upper
BAJOBS	CPUHRB	1917.6	966.0	2869.0
CPUHRB	CORHS	43.7	25.1	62.3
CPUHRT	LOGHRS	7.5	3.3	11.7
COHRS	TAPEHRS	5027.0	3213.0	6840.0
TAPEHRS	COHRS	286.8	127.3	446.2
LOGHRS	LOGONS	1580.7	1358.0	1803.0
LOGONS	LOGHRS	3495.8	3001.0	3990.0
LNCARDIN	TAPEHRS	1413.7	1027.0	1800.0

The 90% confidence interval for the conditional mean (regression line) and the dependent variable are given in Figure 6 for LOGHRS versus LOGONS. The ratio of $s_{y.x}$ to s_y for these two variables was the smallest for all the SLR's done. However, Figure 6 adequately demonstrates that the confidence interval is still rather large and a reduction in the interval was necessary. This was achieved using MLR.

Multiple Linear Regression

The MLR results indicated the shorter time-series was even more appropriate. For the longer time-series, only five of the regressions gave an R^2 greater than 0.9. Some of the standard error of the estimates were 10 times larger than those achieved for the shorter time-series. Therefore, the following discussion will be restricted to the shorter time-series; June 1976 to March 1977.

A summary of the MLR results[8] follows:

Dependent Variable	R^2	$s_{y.x_1x_2...x_n}$	s_y
BAJOBS	0.98896	139.9	627.7
CPUHRB	0.99875	1.9	25.8
CPUHRT	0.98945	0.5	2.2
COHRS	0.99901	232.2	3478.8
TAPEHRS	0.97745	61.6	306.0
LOGHRS	0.99965	14.4	369.1
LOGONS	0.99954	37.3	822.2
LNCARDIN	0.99572	41.2	297.4

All regressions were significant at the 0.05 level of significance. The critical value for the F test (F_c) and the F statistic (F_s) are:

Dependent Variable	F_s	p-1	n-p	F_c
BAJOBS	25.58	7	2	19.30
CPUHRB	228.54	7	2	19.30
CPUHRT	26.81	7	2	19.30
COHRS	288.32	7	2	19.30
TAPEHRS	54.19	4	5	5.05
LOGHRS	827.01	7	2	19.30
LOGONS	623.35	7	2	19.30
LNCARDIN	66.52	7	2	19.30

The variables LOGHRS, LOGONS and LNCARDIN were not introduced into the equation for TAPEHRS. Hence, the different measure for the critical value of the F test. The regression for TAPEHRS was considered satisfactory without the introduction of these variables. Therefore, they were not included in the regression equation. All three variables failed the F ratio test of statistical significance.

---

[8] A comprehensive summary of the MLR results is given in Appendix F.

To allow statistical inference to be drawn concerning the appropriateness of the regression equation, it is necessary that the distribution of residuals about the regression line be normally distributed with a constant variance for all values of the independent variable. SPSS provides a plot of the standardized residual versus the standardized dependent variable to show the distribution of these residuals and allow a judgement to be made concerning the distribution of the residuals. Figure 8 shows the residual plot for TAPEHRS. Since TAPEHRS had the lowest coefficient of determination, the residuals for all other variables were more closely distributed around the regression line.

To determine the consistency of the MLR results, an attempt was made to find a solution to the series of regression equations. To do this, the equations were expressed in matrix form as follows:

$$\underline{z} = [\underline{M}]\underline{z} + \underline{c}$$

where $\underline{z}$ is the column vector of variables in the equations.
$\underline{M}$ is the matrix of estimated values of the coefficients in the regression equation. All diagonal coefficients will be equal to zero.
$\underline{c}$ is the column vector of constants in the regression equations.

Using I as the identity matrix, this may be expressed as:

$$[I]\underline{z} = [\underline{M}]\underline{z} + \underline{c}$$
$$[\underline{M}-I]\underline{z} = -\underline{c}$$
$$[\underline{M}']\underline{z} = \underline{c}^T$$

where $\underline{M}'$ is matrix $\underline{M}$ with minus ones as the diagonal coefficients.
$\underline{c}'$ is the negative vector of vector $\underline{c}$.

```
         -2.0     -1.0       0.0      1.0       2.0
       .YX+---------+---------+---------+---------+XY.
        Y                     I                     Y
        X                     I                     X
S   1.0 +                     I                     +
T       I                     I                     I
A       I                     I                     I
N       I                     I                     I
D       I                     I                     I
A       I                     I                     I
R       I                     I                     I
D       I                     I                     I
I       I               *  *  I                     I
Z       I                     I                   * I
E   0.0 +---------------*2-*--I---------------------+
D       I                 *   I                     I
        I                    *I      *              I
R       I                     I                     I
E       I                     I                     I
S       I                     I                     I
I       I                     I                     I
D       I                     I                     I
U       I                     I                     I
A       I                     I                     I
L  -1.0 +                     I                     +
        X                     I                     X
        Y                     I                     Y
       .YX+---------+---------+---------+---------+XY.
         -2.0     -1.0       0.0      1.0       2.0
```

PREDICTED STANDARDIZED DEPENDENT VARIABLE

FIGURE 8. RESIDUAL PLOT FOR MLR RESULTS--TAPEHRS--AFIT/CREATE--JUNE 1976 TO MARCH 1977

Using the method of Gaussian elimination, an attempt was made to solve several of the sets of regression equations calculated in this research. The solutions consistently gave negative values for some variables. This was inappropriate. The method of Linear Programming (LP) was used to constrain the variables to a positive solution using a "contrived" objective function.

The objective function was expressed in the form:

$$\text{Minimize} \quad Z = \sum k_i z_i \quad \ldots\ldots\ldots\ldots \quad (4.1)$$

$$\text{Subject To:} \quad [\underline{M}']\underline{z} = \underline{c}'$$

where k_i is the coefficient of the ith term; a value of plus one was used.

z_i is the ith variable in vector $\underline{z}$.

The results of the LP formulation for the June 1976 to March 1977 time-series for eight and seven variables (LNCARDIN excluded) compared to the means of the data are:

	LP Solution 8 Variables	LP Solution 7 Variables	Mean
BAJOBS	2225.34	1916.81	1917.60
CPUHRB	68.90	43.72	43.70
CPUHRT	8.30	7.50	7.50
COHRS	8417.75	5032.80	5027.00
TAPEHRS	575.43	287.37	286.80
LOGHRS	1648.34	1582.14	1580.70
LOGONS	3509.70	3498.14	3495.80
LNCARDIN	1219.53	-	1413.70

Changing the coefficients of the objective function variables (k_i in equation 4.1) does not change the solution

[9] The system of linear equations, objective function, and linear program output for 7 variables are given in Table 3.

TABLE 3. LINEAR PROGRAMMING FORMULATION AND SOLUTION-- AFIT/CREATE--APRIL 1976 TO MARCH 1977

```
CONSTRAINTS:

C1:-1.0X1+33.4993X2-36.6039X3+.04843X4-2.07815X5-1.17811X6
     +.62789X7=-747.73375;

C2:.01996X1-1.0X2+.82322X3+.00229X4+.03538X5-.00166X7
     =16.59985;

C3:-.00418X1+.15819X2-1.0X3-.00052X4-.00313X5+.00702X6
     -.00189X7=-7.61797;

C4:.53251X1+42.44436X2-49.9604X3-1.0X4+6.61776X5+4.68713X6
     -2.02039X7=-281.29883;

C5:-.19782X1+6.53576X2-4.48513X3+.05557X4-1.0X5=-134.7805;

C6:-.13593X1+7.03272X3+.04892X4-.23734X5-1.0X6+.45563X7
     =-18.09162;

C7:.30989X1-10.10999X3-.10246X4+.4578X5+2.13915X6-1.0X7
     =20.375;

OBJECTIVE:AFIT/CREATE:X1+X+X3+X4+X5+X6+2X7;
MINIMIZE:AFIT/CREATE;


************************
** SIMPLEX SOLUTION **
************************

THE PROBLEM IS FEASIBLE

NUMBER OF ITERATIONS =    9

OPTIMAL VALUE FOR AFIT/CREATE =        12368.48
```

VARIABLE		VALUE	EFFECT ON OBJECTIVE FUNCTION
X1	=	1916.81	0.00
X2	=	43.72	0.00
X3	=	7.50	-0.00
X4	=	5032.80	-0.00
X5	=	287.37	0.00
X6	=	1582.14	0.00
X7	=	3498.14	-0.00

to the series of equations, only the value of the objective function. That is, there is only one point in the eight (or seven) dimensions where the system of linear equations meet. With the required accuracy, the solution should be the mean values of the variables included in the series of equations. The solution for the seven variables in Table 3 indicates this result.

The variability of the solution for eight variables can be attributed to imprecision in the coefficients of the independent variables. A small change in only one coefficient can have a significant effect on the solution. In the system of linear equations obtained for the total CREATE system, the coefficient for BAJOBS as an independent variable in the equation for CPUHRT was -0.00077. Other coefficients were as high as 40.57624. The coefficient was changed to -0.00079, resulting in the following change to the solution:

	Coefficient of BAJOBS -0.00077	-0.00079	Mean
BAJOBS	12483.60	12383.64	12419.42
CPUHRB	430.57	428.53	428.34
CPUHRT	82.04	80.77	81.89
COHRS	37239.11	37085.45	37015.36
TAPEHRS	1826.61	1832.72	1814.56
LOGHRS	7296.56	7189.16	7252.72
LOGONS	14805.53	14615.27	14716.00

The effect of the small change in a single coefficient is evident. The conclusion that can be drawn: SPSS provides for five decimal place accuracy. This is not sufficient when coefficients are of the order of 10^{-4}.

Trend Analysis

Regressing each variable against time produced a SLR model which indicated the trend of the variable against time. The results of the regression are shown in Figure 9 (pages 66 to 69) using the time-series April 1975 to March 1977 as the base series. BAJOBS, CPUHRT, LOGHRS and LOGONS indicate a decrease, while CPUHRB, COHRS, TAPEHRS and LNCARDIN show an increase.

Validation of Research Methodology

The techniques applied to the AFIT/CREATE data series were applied to the data base available for CREATE[10] as a whole in order to determine if the methodology used was appropriate to another series of data. LNCARDIN was not included in the CREATE analysis since data for this variable were not readily available from the monthly accounting records.

In general, the results achieved for the AFIT/CREATE time-series were replicated using the CREATE time-series. Specifically, some of the results achieved were:

1. There is less variability in the use of CREATE as a whole compared to AFIT/CREATE.

2. There is a declining trend in the use of all resources except COHRS.

---

[10]Graphs for the CREATE data, and comprehensive research results are contained in Appendix G.

3. The shorter the time-series, the higher the bivariate correlation coefficients achieved. For the April 1976 to March 1977[11] time-series, seven coefficients were above 0.8, and of those, two were above 0.9.

4. SLR analysis provided two regressions with an R^2 above 0.8.

5. The MLR results for the last 12 months of data gave far better results than those for the 39 month or 24 month time-series. All of the coefficients of determination were higher, and all of the standard errors of the estimate were smaller (except one) for the 12 month time-series.

6. The TCAST results achieved were generally better than the results for the AFIT/CREATE data-series. TCAST selected a cycle of one for four of the variables. In comparison to the mean of the data, the MAD was generally smaller than for the AFIT/CREATE series. This allowed smaller confidence intervals to be calculated and was attributed to less variability in the data-series.

---

[11]The 12 month data-series was chosen to give a complete 12 months of data whereas for the AFIT/CREATE analysis, only 10 months of data were used because of the introduction of SPSS.

CHAPTER V

CONCLUSIONS AND RECOMENDATIONS

Conclusions

General

Some general observations concerning the characteristics of the data base should be highlighted.

1. There is high variability in the parameters representing AFIT/CREATE usage. CPUHRB, COHRS and TAPEHRS have the highest variability. In general, the peaks in computer workload for AFIT/CREATE occur just prior to the graduation of one of the Graduate Logistics classes. Therefore, the conclusion by Anderson and Purnell that ". . . heavy usage rates are a function of the heavy use of large programs to complete thesis work [1:61]," was supported. However, the high variability of the data contributed to the inability of TCAST to give an accurate forecasting model.

2. The trend analysis of the data indicated some decrease and some increase in computer usage. This is contrary to the interview results obtained Anderson and Purnell which indicated a 20% increase for both time-sharing and batch (1:54). Also, some results, while accurate, are not necessarily "intuitively obvious". That is, while CPUHRB

is showing an increasing trend, BAJOBS is actually showing a decreasing trend. A positive correlation would generally be considered applicable rather than a negative correlation.

Length of Time Series

The most significant observation of this research was that use of a long data series can be detrimental when attempting to define an accurate forecasting model. Both the time-series forecasting and regression analysis provided more accurate forecasts using a short data series.

The reason for the phenomenon is the change in the true process with time (non-stationary). Two significant changes in the AFIT/CREATE usage pattern between January 1974 and March 1977 were the removal of GASP4B FORTRAN based time-sharing simulation from the curriculum with the 1975 graduating classes and the introduction of SPSS with the 1977 "A" Class.

TCAST Results

In general, the conclusion by Anderson and Purnell (1:61) that the change in schedule of the Graduate Logistics Classes in 1975 contributed to TCAST experiencing difficulty in obtaining a "good" fit to the actual data points was confirmed. Smaller MAD errors were obtained by removing the data from the time-series that applied to the period before the change. With the exception of TAPEHRS, it was concluded that TCAST gave acceptable forecast models.

However, it was concluded that the dominant cycle selected by TCAST may not necessarily be the cycle that will provide the most accurate model. It is appropriate to compute forecast models with different cycles, choosing those cycles that give a cyclic error relatively close to the cycle selected by TCAST. The model selected should have as small a MAD as possible, and have a composite that is representative of the actual data.

The forecasts obtained using a lead-time of 12 months are given in Figures 9A through 9G for the April 1975 to March 1977 data-series. For aggregate planning purposes, the trend line could be used for forecasting. However, for highly variable data, a large standard error of the estimate would result, giving a large confidence interval. For specific forecast estimates, TCAST can provide an acceptable model as demonstrated in Figure 5 for LOGHRS. TCAST is unable to provide an acceptable model if a high end data value is encountered as for TAPEHRS or the process is changing significantly over a period of time. Brown's adaptive smoothing technique appears appropriate in this case (5:168). Instead of estimating the coefficients in the model with a fixed origin in time, the coefficients in a model should be re-estimated but with time relative to the most recent observation and the oldest data value removed.

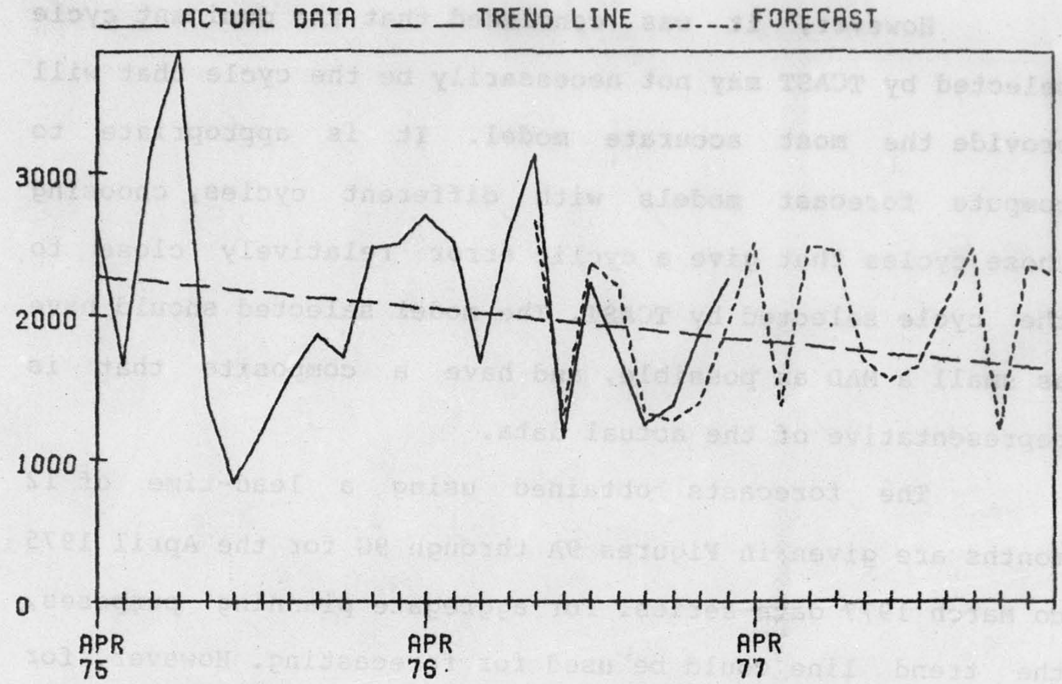


Figure 9A. GRAPH: BAJOBS, FORECAST AND TREND

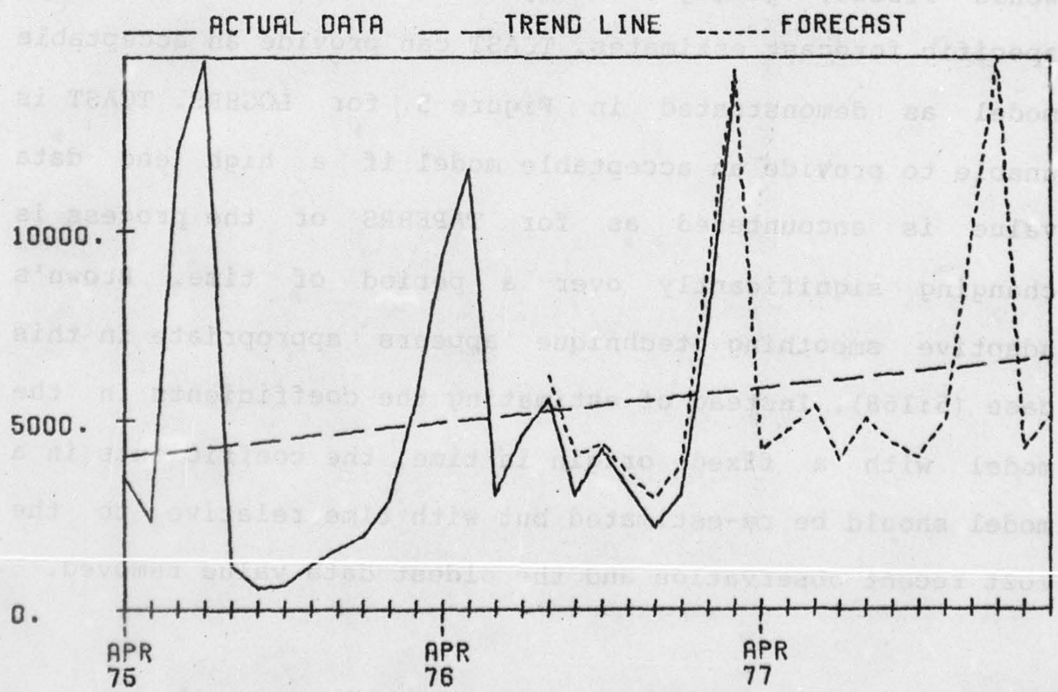


Figure 9B. GRAPH: COHRS, FORECAST AND TREND

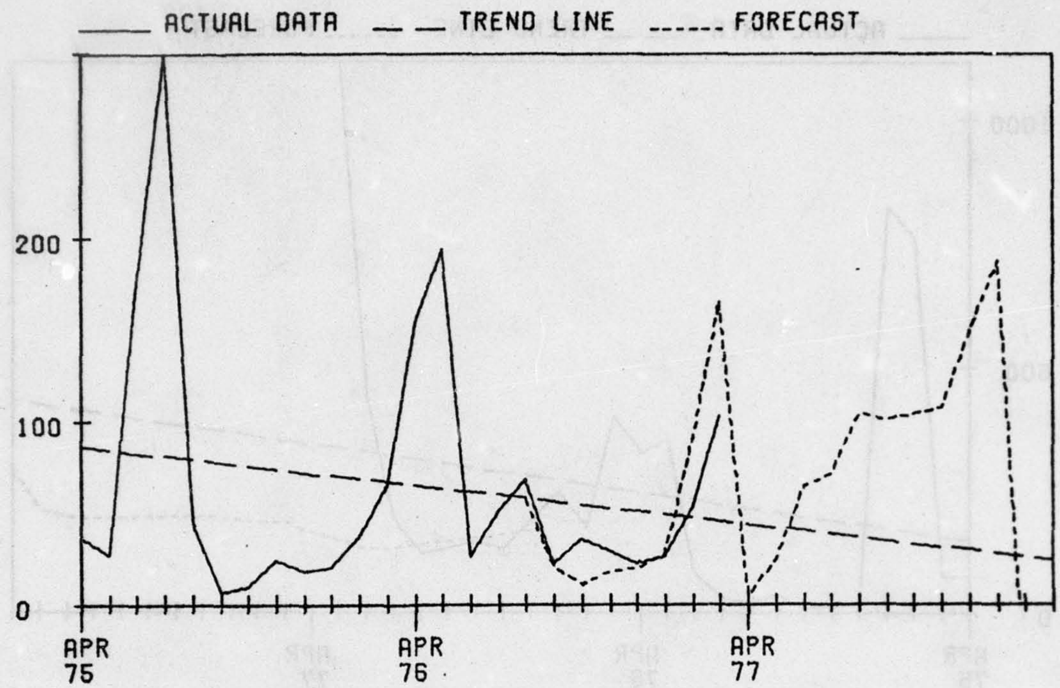


Figure 9C. GRAPH: CPUHRB, FORECAST AND TREND

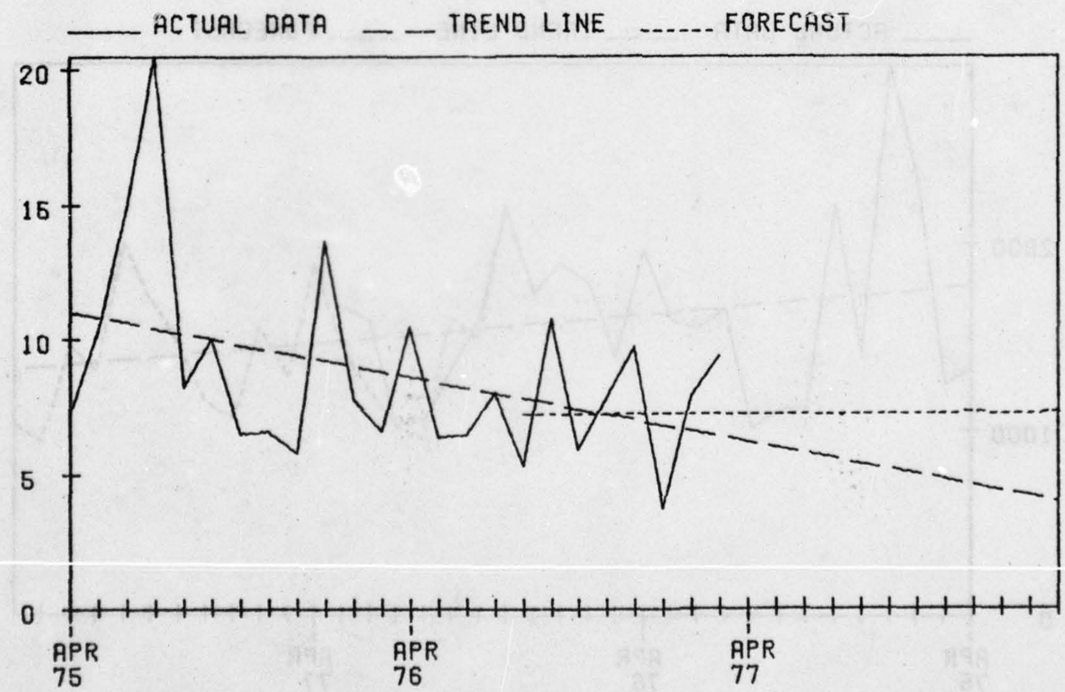


Figure 9D. GRAPH: CPUHRT, FORECAST AND TREND

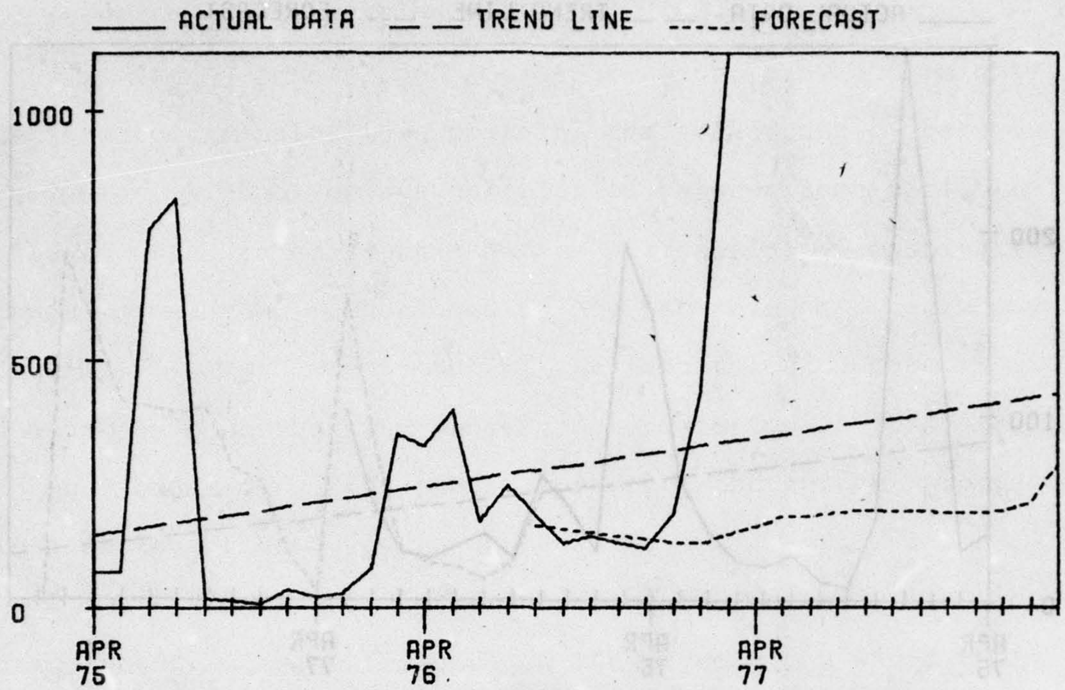


Figure 9E. GRAPH: TAPEHRS, FORECAST AND TREND

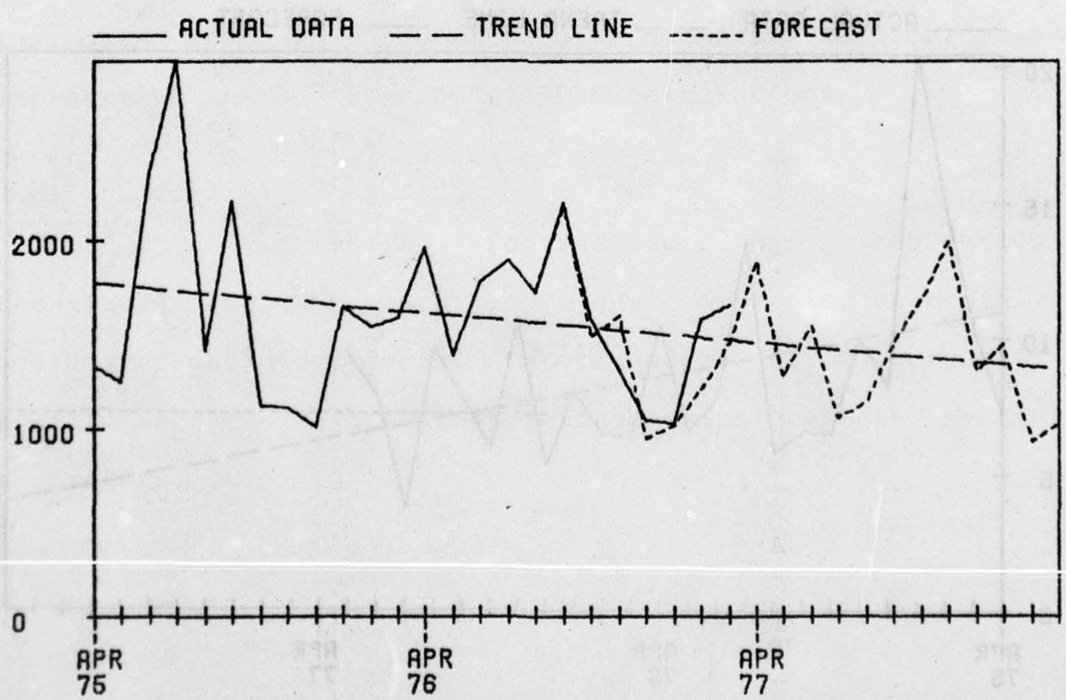


Figure 9F. GRAPH: LOGHRS, FORECAST AND TREND

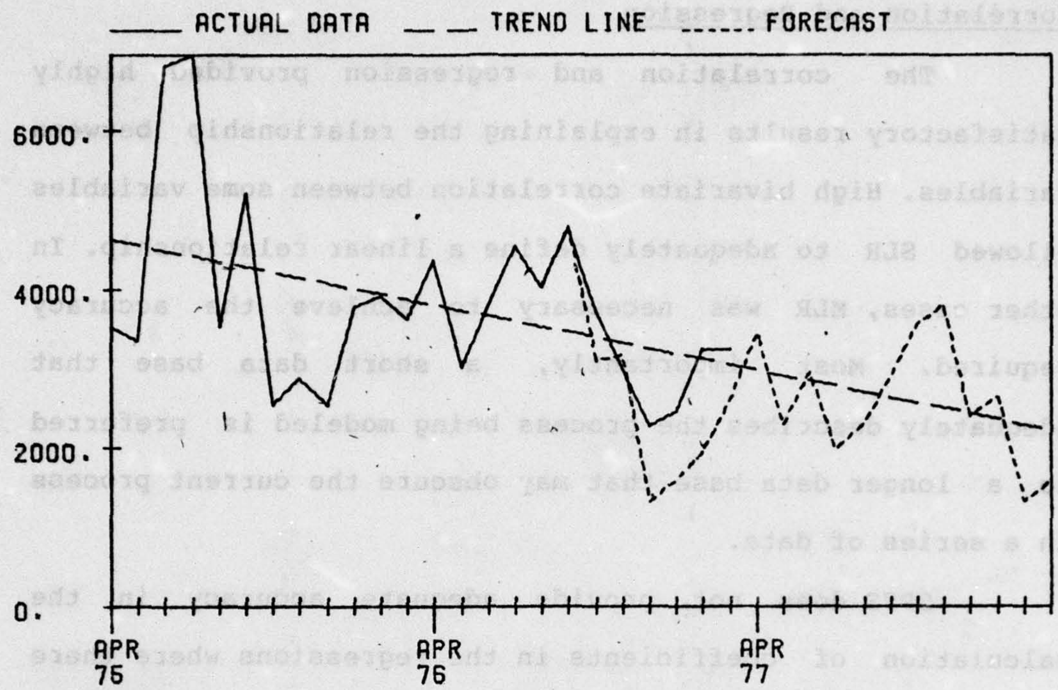


Figure 9G. GRAPH: LOGONS, FORECAST AND TREND

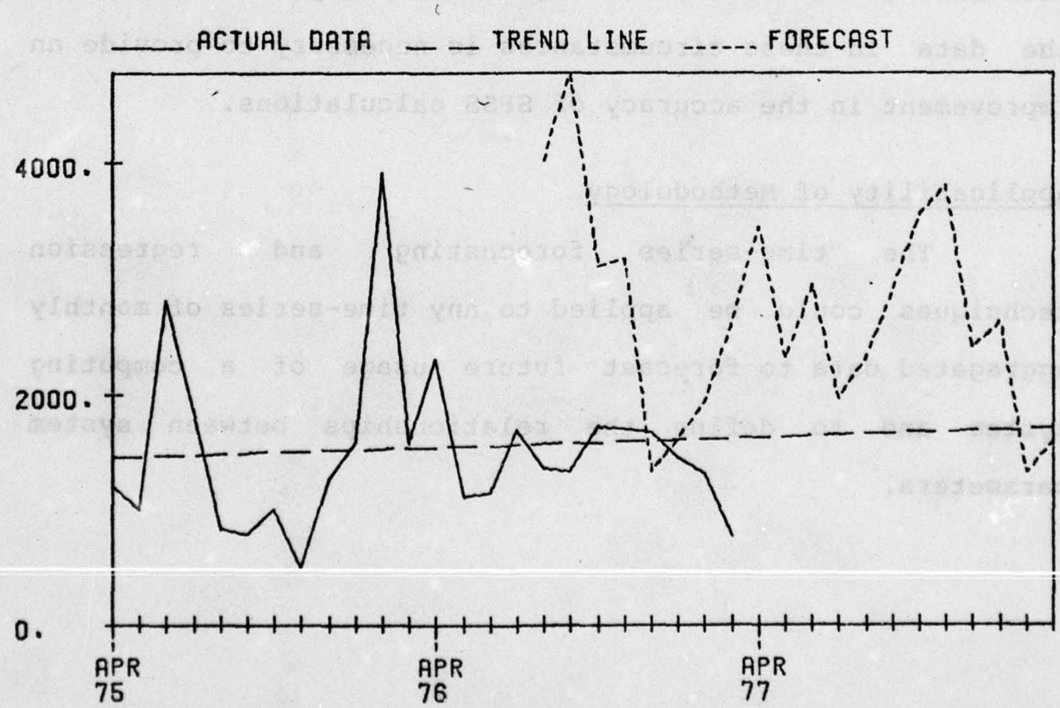


Figure 9H. GRAPH: LNCARDIN, FORECAST AND TREND

Correlation and Regression

The correlation and regression provided highly satisfactory results in explaining the relationship between variables. High bivariate correlation between some variables allowed SLR to adequately define a linear relationship. In other cases, MLR was necessary to achieve the accuracy required. Most importantly, a short data base that adequately describes the process being modeled is preferred to a longer data base that may obscure the current process in a series of data.

SPSS does not provide adequate accuracy in the calculation of coefficients in the regressions where there is a significant difference in the magnitude of data for different variables used in regression analysis. Normalizing the data in these circumstances is necessary to provide an improvement in the accuracy of SPSS calculations.

Applicability of Methodology

The time-series forecasting and regression techniques could be applied to any time-series of monthly aggregated data to forecast future usage of a computing system and to define the relationships between system parameters.

Answers to Research Questions

Can the Anderson and Purnell model be further developed for use as an accurate forecasting tool for AFIT/CREATE support requirements? The use of time-series forecasting techniques as proposed by Anderson and Purnell is adequate for a forecasting model. Sufficient accuracy can be obtained using time-series forecasting without introducing regression analysis.

Does the model have a wider application for use in future computer system specifications? The results obtained for the AFIT/CREATE system were adequately validated using data from the total CREATE system. The regression analysis adequately defined the characteristics of use of the system by providing a relationship between variables. Using an estimate or prediction for each variable, new, synchronized system requirements could be established using the regression equations.

Can the model be used to define separate system requirements for AFIT? The recent relocation of the Graduate Logistics School near the Engineering School provides the impetus to consider a separate "stand alone" computer system for the AFIT schools in the same manner used at larger universities. The parameter averages used in the regression tend to give baseline "requirements" for such a system applicable to Graduate Logistics. Extrapolating these "requirements" through use of a new parameter such as

student enrollment could provide the baseline for a separate AFIT computer system.

Recommendations

Techniques

It is felt that both time-series forecasting and linear regression techniques should be applied to the monthly aggregated usage data to allow management to forecast system usage and define the characteristics of system use. The techniques provide a sufficiently accurate and simple model that appropriately defines the system being used. For systems that have the necessary statistical packages available, the cost of applying the techniques on a monthly basis would be minimal.

The graphic features within this research increase the power of explanation of many of the aspects of this research. This applies in particular, though not exclusively, to the monthly aggregated data. Managers and designers of computing systems should consider incorporating appropriate visual displays of the most frequently sought information such as the growth and cyclic nature of system usage into accounting routines.

Application For Further Study

Methodology. The use of enrollments and student population were not pursued in this research. The effect of these variables on the use of the AFIT/CREATE system has not

been adequately defined. Further research could be conducted in this area to determine if student population does affect system usage or whether the level of system usage is dependent upon the characteristics of use.

Data. The data used in this research was monthly-aggregated data and does not provide an indication of the daily use of the computing system or the characteristics of its use except on a monthly basis. The techniques applied by Hunt, et. al. (7) and the use of probability models appear to be appropriate for research applied to characteristics of system use.

Currently, some data of each individual log-on record is "lost" when the monthly CREATE accounting report is compiled. However, AFIT Data Automation staff are attempting to retain these daily log-on records on tape. This would allow research using probability models to be attempted.

APPENDIXES

APPENDIX A

LIST OF VARIABLES AND
THEIR MNEMONICS

APPENDIX A

LIST OF VARIABLES AND THEIR MNEMONICS

MNEMONIC	VARIABLE	
BAJOBS	Batch Jobs	
COHRS	Core Hours	
COHRSP	Core Hours	Prime
CORHSNP	Core Hours	Non-Prime
CPUHRB	CPU Hours Batch	
CPUHRBP	CPU Hours Batch	Prime
CPUHRBNP	CPU Hours Batch	Non-Prime
CPUHRT	CPU Hours Time-Sharing	
CPUHRTP	CPU Hours Time-Sharing	Prime
CPUHRTNP	CPU Hours Time-Sharing	Non-Prime
LOGHRS	Log-On Hours	
LOGHRSP	Log-On Hours	Prime
LOGHRSNP	Log-On Hours	Non-Prime
LOGONS	Number of Log-Ons	
LNCARDIN	Lines CARDIN	

APPENDIX B

AGGREGATED DATA FOR AFIT/CREATE USAGE-- JANUARY 1974 TO MARCH 1977

APPENDIX B

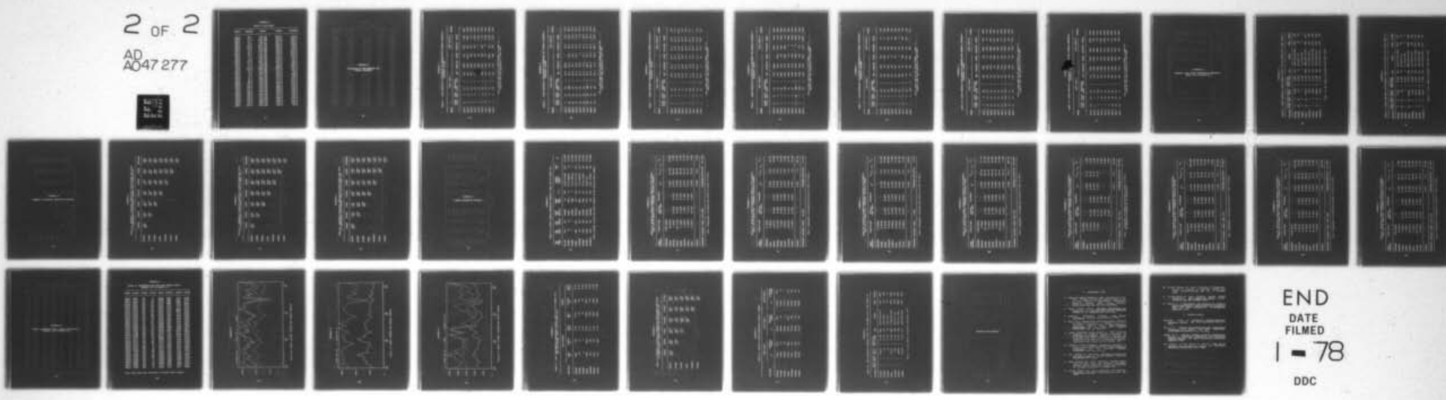
TABLE 4. AGGREGATED DATA FOR AFIT/CREATE USAGE-- JANUARY 1974 TO MARCH 1977

MONTH	BAJOBS	CPUHRB	CPUHRT	COHRS
JAN74	5036.0	26.3624	6.2603	2904.0570
FEB74	1740.0	18.1170	21.8710	2197.5067
MAR74	1863.0	16.1584	11.6515	1956.5480
APR74	1912.0	20.4529	18.4038	2476.0778
MAY74	2283.0	21.6943	13.6425	2188.6737
JUN74	2615.0	82.1925	55.0941	8187.2261
JUL74	2664.0	81.6403	113.5946	11218.9711
AUG74	1996.0	23.4888	22.2854	2504.3466
SEP74	2724.0	24.6737	34.4287	3448.9601
OCT74	1776.0	16.1220	17.1741	1489.6019
NOV74	3244.0	58.0126	12.5973	4704.7570
DEC74	2273.0	88.0343	11.8717	4911.9182
JAN75	1564.0	21.7940	40.1670	3321.2786
FEB75	1786.0	14.8607	38.9452	2622.9035
MAR75	2223.0	33.2802	16.0222	3196.7436
APR75	2628.0	36.9185	7.2149	3539.9496
MAY75	1659.0	27.4383	10.3188	2352.8635
JUN75	3214.0	172.5462	14.9211	12136.6229
JUL75	3844.0	299.4874	20.4711	14513.2323
AUG75	1391.0	44.9589	8.1877	1198.0031
SEP75	829.0	7.2480	10.0349	545.6394
OCT75	1189.0	9.2264	6.5139	664.6823
NOV75	1564.0	23.8366	6.5299	1201.9797
DEC75	1870.0	18.8958	5.7312	1606.6381
JAN76	1706.0	20.5460	13.6507	1912.4603
FEB76	2470.0	36.4945	7.7615	2870.0729
MAR76	2504.0	65.1385	6.5459	5515.0746
APR76	2705.0	153.9560	10.4486	9324.8970
MAY76	2514.0	193.3324	6.3885	11606.2728
JUN76	1671.0	26.0495	6.4399	3026.6300
JUL76	2567.0	51.7833	8.0288	4612.5805
AUG76	3120.0	68.1732	5.2830	5554.7325
SEP76	1150.0	22.0228	10.7449	3022.5149
OCT76	2232.0	35.5314	5.9027	4257.8051
NOV76	1725.0	29.0282	7.8040	3203.8237
DEC76	1229.0	22.5251	9.7053	2149.8423
JAN77	1368.0	26.3806	3.7205	2999.0845
FEB77	1860.0	53.3490	7.9005	7615.4727
MAR77	2254.0	102.1224	9.3660	13827.1530

AD-A047 277 AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OHIO SCHO--ETC F/G 5/1
A MODEL FOR FORECASTING THE COMPUTER REQUIREMENTS FOR THE AIR F--ETC(U)
SEP 77 H H DETJEN, D C JOHNSTON

UNCLASSIFIED AFIT-LSSR-1-77B NL

2 OF 2
AD
A047 277



END
DATE
FILMED
1 - 78
DDC

APPENDIX B

TABLE 4.(continued)

MONTH	TAPEHRS	LOGHRS	LOGONS	LNCARDIN
JAN74	20.7	790.170	4076.0	463.0
FEB74	42.5	1957.400	4614.0	1573.0
MAR74	18.5	1683.220	3841.0	1454.0
APR74	11.1	1570.521	3657.0	1607.0
MAY74	25.5	1392.780	3231.0	1764.0
JUN74	29.8	2013.650	4054.0	1939.0
JUL74	203.9	2636.570	4538.0	758.0
AUG74	23.4	2415.600	6105.0	1656.0
SEP74	11.8	2190.700	5786.0	2214.0
OCT74	20.7	1237.446	3240.0	1656.0
NOV74	26.1	1357.296	3734.0	2252.0
DEC74	8.5	1308.950	3119.0	2126.0
JAN75	11.8	1570.640	3348.0	1141.0
FEB75	26.0	1507.950	3362.0	914.0
MAR75	112.6	2083.920	5025.0	990.0
APR75	71.3	1328.000	3526.0	1208.0
MAY75	71.4	1249.380	3350.0	1004.0
JUN75	761.8	2368.560	6803.0	2817.0
JUL75	824.0	2959.200	6936.0	1828.0
AUG75	17.7	1421.340	3521.0	834.0
SEP75	14.0	2209.890	5207.0	792.0
OCT75	8.1	1122.570	2530.0	1000.0
NOV75	37.7	1114.460	2875.0	499.0
DEC75	21.1	1006.370	2538.0	1260.0
JAN76	30.0	1648.030	3773.0	1578.0
FEB76	80.8	1541.450	3915.0	3915.0
MAR76	351.6	1586.960	3606.0	1576.0
APR76	326.4	1967.030	4352.0	2304.0
MAY76	397.9	1392.140	3008.0	1110.0
JUN76	174.3	1778.530	3800.0	1149.0
JUL76	246.6	1898.590	4523.0	1680.0
AUG76	189.1	1727.370	4026.0	1357.0
SEP76	130.0	2202.640	4753.0	1335.0
OCT76	143.2	1581.530	3703.0	1709.0
NOV76	130.8	1311.040	2982.0	1692.0
DEC76	118.5	1040.550	2260.0	1673.0
JAN77	186.4	1026.520	2446.0	2446.0
FEB77	431.4	1583.120	3237.0	1311.0
MAR77	1117.4	1657.440	3228.0	780.0

APPENDIX C

VALIDATION OF THE ANDERSON AND PURNELL RESEARCH

APPENDIX C

TABLE 5. TCAST ANALYSIS--VALIDATION OF ANDERSON AND PURNELL RESEARCH--VARIABLE: BAJOBS

MONTH	LEAD TIME=2							LEAD TIME=12*		ACTUAL
	TCAST CYCLE	CYCLE USED	SMOOTHING ALPHA	SMOOTHING TYPE	MAD	FORECAST	ERROR	FORECAST	ERROR	
JUL76	5	5	.17	1	623	2050	-517	1101	-1466	2567
AUG76	5	5	.35	1	621	1439	-1681	287	-2833	3120
SEP76	5	5	.31	1	555	1898	748	373	-777	1150
OCT76	5	5	.30	2	4[illegible]	2748	516	476	-1756	2232
NOV76	5	5	.16	1	[illegible]	1860	135	3	-1722	1725
DEC76	5	5	.29	1	580	2167	938	454	-775	1229
JAN77	5	5	.30	1	557	1615	247	0	-1368	1368
FEB77	5	5	.31	1	612	802	-1058	162	-1698	1860
MAR77	5	5	.30	1	541	1374	-880	478	-1776	2254

*For a lead time of 12: Cycle--TCAST 5, Used 5; Alpha 0.140; Type Smoothing 1; MAD 1344.

APPENDIX C

TABLE 6. TCAST ANALYSIS--VALIDATION OF ANDERSON AND PURNELL RESEARCH--VARIABLE: CPUHRB

MONTH	LEAD TIME=2							LEAD TIME=12*		ACTUAL
	TCAST CYCLE	CYCLE USED	SMOOTHING ALPHA	SMOOTHING TYPE	MAD	FORECAST	ERROR	FORECAST	ERROR	
JUL76	1	12	.46	3	35.6	448.0	-	42.9	-8.9	51.8
AUG76	10	10	.008	3	26.2	46.1	-22.1	50.2	-18.0	68.2
SEP76	10	10	.03	3	28.4	47.7	25.7	38.2	16.2	22.0
OCT76	10	10	.03	3	27.3	49.2	13.7	32.7	-2.8	35.5
NOV76	10	10	.01	3	26.2	29.8	.8	37.8	8.8	29.0
DEC76	10	10	.01	3	25.3	74.6	52.1	61.1	38.6	22.5
JAN77	10	10	.001	1	25.5	80.4	54.1	72.1	45.8	26.3
FEB77	10	10	.10	1	29.6	103.9	50.6	136.7	83.4	53.3
MAR77	10	10	.09	1	30.3	13.3	-88.8	188.3	86.2	102.1

*For a lead time of 12: Cycle--TCAST 10, Used 10; Alpha 0.020; Type Smoothing 3; MAD 7043.

APPENDIX C

TABLE 7. TCAST ANALYSIS--VALIDATION OF ANDERSON AND PURNELL RESEARCH--VARIABLE: CPUHRT

MONTH	LEAD TIME=2							LEAD TIME=12*		ACTUAL
	TCAST CYCLE	CYCLE USED	SMOOTHING ALPHA	SMOOTHING TYPE	MAD	FORECAST	ERROR	FORECAST	ERROR	
JUL76	1	6	.7Ø	1	19.Ø	44.5	36.5	84.4	76.8	8.Ø
AUG76	1	6	.7Ø	1	18.6	24.4	19.1	73.6	68.3	5.3
SEP76	1	6	.7Ø	1	19.Ø	Ø.7	-1Ø.Ø	69.3	58.6	1Ø.7
OCT76	7	7	.13	1	18.3	16.9	11.Ø	68.1	62.2	5.9
NOV76	7	7	.21	1	17.4	38.9	31.9	65.8	58.Ø	7.8
DEC76	7	7	.Ø9	1	17.7	Ø.Ø	-9.7	74.4	64.7	9.7
JAN77	7	7	.Ø7	2	19.Ø	Ø.Ø	-3.7	99.7	96.Ø	3.7
FEB77	1	6	.32	2	18.4	2Ø.Ø	12.1	88.9	81.Ø	7.9
MAR77	1	6	.6Ø	1	18.Ø	2.7	-6.7	84.6	75.2	9.4

*For a lead time of 12: Cycle--TCAST 1, Used 6; Alpha Ø.ØØ1; Type Smoothing 1; MAD Ø576.

APPENDIX C

TABLE 8. TCAST ANALYSIS--VALIDATION OF ANDERSON AND PURNELL RESEARCH--VARIABLE: COHRS

MONTH	LEAD TIME=2							LEAD TIME=12*		ACTUAL
	TCAST CYCLE	CYCLE USED	SMOOTHING ALPHA	SMOOTHING TYPE	MAD	FORECAST	ERROR	FORECAST	ERROR	
JUL76	12	12	.61	2	1356	24690	-	11227	6615	4612
AUG76	1	2	.10	1	4319	2883	-2671	9273	3719	5554
SEP76	1	2	.08	1	4005	3914	892	7106	4084	3022
OCT76	1	2	.001	1	3556	3170	-1087	4895	638	4257
NOV76	1	2	.16	1	3940	3879	676	0	-3203	3203
DEC76	1	2	.07	1	3568	3378	1184	0	-2194	2194
JAN77	1	10	.009	3	2140	8427	5428	0	-2999	2999
FEB77	1	5	.001	1	3208	6630	-985	227	-7388	7615
MAR77	1	5	.02	3	3296	6952	-6877	3152	-10677	13829

*For a lead time of 12: Cycle--TCAST 1, Used 5; Alpha 0.100; Type Smoothing 3; MAD 1808.

APPENDIX C

TABLE 9. TCAST ANALYSIS--VALIDATION OF ANDERSON AND PURNELL RESEARCH--VARIABLE: TAPEHRS

MONTH	LEAD TIME=2							LEAD TIME=12*		ACTUAL
	TCAST CYCLE	CYCLE USED	SMOOTHING ALPHA	SMOOTHING TYPE	MAD	FORECAST	ERROR	FORECAST	ERROR	
JUL76	4	4	.60	1	158	562	315	339	92	247
AUG76	1	4	.01	1	123	85	-104	184	-5	189
SEP76	1	4	.01	3	118	109	-21	184	314	130
OCT76	1	4	.004	3	118	236	93	198	55	143
NOV76	4	4	.02	3	107	267	137	260	130	130
DEC76	4	4	.07	1	103	127	9	116	-2	118
JAN77	4	4	.12	1	109	96	-90	126	-60	186
FEB77	1	4	.02	2	98	156	-275	155	-276	431
MAR77	4	4	.001	1	96	266	-851	260	-	1117

*For a lead time of 12: Cycle--TCAST 1, Used 4; Alpha 0.120; Type Smoothing 1; MAD 5824.

APPENDIX C

TABLE 10. TCAST ANALYSIS--VALIDATION OF ANDERSON AND PURNELL RESEARCH--VARIABLE: LOGHRS

MONTH	LEAD TIME=2							LEAD TIME=12*		ACTUAL
	TCAST CYCLE	CYCLE USED	SMOOTHING ALPHA	SMOOTHING TYPE	MAD	FORECAST	ERROR	FORECAST	ERROR	
JUL76	6	6	.60	1	388	2382	432	2694	795	1899
AUG76	12	12	.45	1	186	1656	-71	1754	27	1727
SEP76	12	12	.53	1	233	1476	-726	2060	-142	2202
OCT76	12	12	.19	3	227	747	-834	1050	-531	1581
NOV76	12	12	.19	3	232	1034	-277	1106	-205	1311
DEC76	12	12	.19	3	257	1383	343	1023	-17	1040
JAN77	12	12	.53	1	246	1663	637	1499	473	1026
FEB77	12	12	.45	1	235	1768	185	1693	110	1583
MAR77	12	12	.40	1	236	1518	-139	1759	102	1657

*For a lead time of 12: Cycle--TCAST 12, Used 12; Alpha 0.150; Type Smoothing 1; MAD 9504.

APPEND[illegible]

TABLE 11. TCAST ANALYSIS--VALIDATION O[illegible] ANDERSON AND PURNELL RESEARCH--
VARIABLE: LOGONS

MONTH	LEAD TIME=2							LEAD TIME=12*		ACTUAL
	TCAST CYCLE	CYCLE USED	SMOOTHING ALPHA	SMOOTHING TYPE	MAD	FORECAST	ERROR	FORECAST	ERROR	
JUL76	10	10	.35	1	637	1940	-2583	5606	1085	4523
AUG76	1	3	.07	1	730	2979	-1047	4325	299	4026
SEP76	1	2	.001	1	815	4416	-337	5005	252	4753
OCT76	1	10	.25	1	692	3288	-415	2358	-1345	3703
NOV76	1	10	.35	1	731	4926	1944	2713	-269	2982
DEC76	1	10	.08	1	700	4326	2066	2225	-35	2260
JAN77	10	10	.07	2	894	3505	1059	3117	671	2446
FEB77	10	10	.05	3	887	5056	1819	3425	188	3237
MAR77	10	10	.17	1	902	3918	690	3432	204	3228

*For a lead time of 12: Cycle--TCAST 1, Used 12; Alpha 0.110; Type Smoothing 2; MAD 9008.

APPENDIX D

RESULTS: TIME SERIES FORECASTING ANALYSIS-- APRIL 1975 TO MARCH 1977

APPENDIX D

TABLE 12. TCAST ANALYSIS--AFIT/CREATE--APRIL 1975 TO MARCH 1977--LEAD TIME= 2

VARIABLE	CYCLE TCAST	CYCLE USED	CYCLIC ERROR	SMOOTHING TYPE	SMOOTHING ALPHA	MAD	TREND	MEAN	STDDEV
BAJOBS	8	8	32222	1	0.001	421	2281.152-18.282t*	2052	628
CPUHRB	10	10	471	2	0.090	26	85.959-1.72t	64	26
CPUHRT	1	7	1	1	0.080	2	11.133-0.192t	9	2
COHRS	10	10	364644	2	0.620	1060	4080.949+71.051t	4969	3479
TAPEHRS	1	4	3459	2	0.580	185	147.225+7.827t	245	306
LOGHRS	7	7	14579	1	0.120	238	1771.741-12.664t	1613	369
LOGONS	7	7	80283	1	0.290	530	4614.98-66.205t	3787	822
LNCARDIN	2	8	45104	1	0.210	408	1447.529+7.054t	1536	297

*Time t is the number of months from April 1975.

APPENDIX D

TABLE 13. TCAST ANALYSIS--AFIT/CREATE--APRIL 1975 TO MARCH 1977--LEAD TIME= 12

VARIABLE	CYCLE TCAST	CYCLE USED	CYCLIC ERROR	SMOOTHING TYPE	SMOOTHING ALPHA	MAD	TREND	MEAN	STDDEV
BAJOBS	8	8	32222	3	0.030	482	2281.152-18.282t*	2052	627
CPUHRB	10	10	130	3	0.100	25	85.959-1.72t	64	25
CPUHRT	1	1	1	2	0.001	2	11.133-0.192t	9	2
COHRS	10	10	364644	1	0.150	1618	4080.949+71.051t	4969	3478
TAPEHRS	1	1	3370	3	0.020	116	147.225+7.827t	245	305
LOGHRS	7	7	14579	3	0.040	304	1771.741-12.664t	1613	369
LOGONS	7	7	80283	2	0.110	769	4614.98-66.205t	3787	822
LNCARDIN	2	4	44522	3	0.100	704	1447.529+7.054t	1537	297

*Time t is the number of months from April 1975.

APPENDIX E

PEARSON'S BIVARIATE CORRELATION RESULTS

APPENDIX E

TABLE 14. PEARSON'S BIVARIATE CORRELATION COEFFICIENTS AND LEVEL OF SIGNIFICENCE--AFIT/CREATE--JANUARY 1974 TO MARCH 1977

	CPUHRB	CPUHRT	COHRS	TAPEHRS	LOGHRS	LOGONS	LNCARDIN
BAJOBS	.4545 .002	.1023 .268	.4749 .001	.2675 .050	.1128 .247	.3622 .012	.2095 .100
CPUHRB		.0741 .327	.8749 .001	.7249 .001	.4666 .001	.4277 .003	.2364 .074
CPUHRT			.2892 .037	-.0695 .337	.1758 .001	.2381 .072	.0861 .301
COHRS				.8482 .001	.4956 .001	.3615 .012	.1505 .180
TAPEHRS					.3663 .011	.2933 .035	.0705 .335
LOGHRS						.8662 .001	.2002 .111
LOGONS							.3756 .042

APPENDIX E

TABLE 15. PEARSON'S BIVARIATE CORRELATION COEFFICIENTS AND LEVEL OF SIGNIFICENCE--AFIT/CREATE--APRIL 1975 TO MARCH 1977

	CPUHRB	CPUHRT	COHRS	TAPEHRS	LOGHRS	LOGONS	LNCARDIN
BAJOBS	.7765 .001	.3829 .032	.7441 .001	.5958 .001	.5071 .001	.5573 .006	.4645 .002
CPUHRB		.6187 .001	.8903 .001	.7358 .001	.6279 .001	.6108 .001	.2569 .113
CPUHRT			.4816 .009	.4717 .010	.7500 .001	.7616 .001	.2597 .110
COHRS				.9310 .001	.5556 .002	.4930 .007	.2188 .152
TAPEHRS					.5173 .005	.4477 .014	.1324 .269
LOGHRS						.9490 .001	.2791 .093
LOGONS							.3731 .036

APPENDIX E

TABLE 16. PEARSON'S BIVARIATE CORRELATION COEFFICIENTS AND LEVEL OF SIGNIFICENCE--AFIT/CREATE--JUNE 1976 TO MARCH 1977

	CPUHRB	CPUHRT	COHRS	TAPEHRS	LOGHRS	LOGONS	LNCARDIN
BAJOBS	.6805	-.3225	.4281	.2497	.2930	.3835	-.0813
	.015	.182	.109	.243	.206	.137	.412
CPUHRB		.0768	.9372	.8642	.2181	.1086	-.6305
		.417	.001	.001	.272	.383	.025
CPUHRT			.2016	.2687	.3633	.2043	-.1531
			.288	.226	.151	.186	.336
COHRS				.9678	.1864	.0062	-.7360
				.001	.303	.493	.008
TAPEHRS					.1040	-.0974	-.7785
					.387	.394	.004
LOGHRS						.9568	-.3132
						.001	.189
LOGONS							-.0823
							.411

APPENDIX F

LINEAR REGRESSION RESULTS

APPENDIX 6

TABLE 17. SIMPLE LINEAR REGRESSION RESULTS--JUNE 1976 TO MARCH 1977

DEP VAR NAME	DEP VAR MEAN	DEP VAR STDDEV	INDEP VAR NAME (X)	INDEP VAR MEAN	$S_{y.x}$	SIMPLE LINEAR MODEL	HALF CI	R^2
BAJOBS	1917	628	CPUHRB	44	488	1193+16.569X	952	0.4632
CPUHRB	44	26	COHRS	5027	10	8.78+.00695X	19	0.8783
CPUHRT	7	2	LOGHRS	1580	2	4.12+.00213X	4	0.1320
COHRS	5027	3479	TAPEHRS	287	929	1871.7+11.0X	813	0.9365
TAPEHRS	287	306	COHRS	5027	82	-141+.08512X	159	0.9365
LOGHRS	1581	369	LOGONS	3496	114	79.26+.4295X	222	0.9155
LOGONS	3496	822	LOGHRS	1581	253	126+2.13X	494	0.9155
LNCARDIN	1414	297	TAPEHRS	288	198	1630-.75652X	386	0.6060

Note: Integer values have been rounded to the nearest whole number.

APPENDIX 6

TABLE 18. MULTIPLE LINEAR REGRESSION RESULTS--AFIT/CREATE--
JUNE 1976 TO MARCH 1977: DEPENDENT VARIABLE--BAJOBS

INDEP VARIABLE	COEFFICIENT	PARTIAL CORR COEFF	TOLERANCE	R^2	$S_{y.x_1 \ldots x_n}$	F_s
CONSTANT	-3301.56839	-	-	-	-	-
CPUHRB	68.89244	-	-	0.4631	487.80	19.090
TAPEHRS	1.96627	-0.9180	0.2532	0.9155	206.90	1.280
CPUHRT	-265.77249	-0.3931	0.8323	0.9286	205.40	8.210
LOGONS	-3.10026	0.4380	0.6977	0.9423	202.30	4.330
LNCARDIN	2.98057	0.5807	0.3065	0.9617	184.10	6.770
LOGHRS	8.09056	0.2170	0.0081	0.9635	207.60	4.840
COHRS	-0.50315	-0.8349	0.0033	0.9890	139.90	4.600
OVERALL				0.9890	139.9	4.600

DEPENDENT VARIABLE: MEAN= 1917.6000 STANDARD DEVIATION= 627.7086

Note: Independent variables are listed in order of entry.

APPENDIX F

TABLE 19. MULTIPLE LINEAR REGRESSION RESULTS--AFIT/CREATE--
JUNE 1976 TO MARCH 1977: DEPENDENT VARIABLE--CPUHRB

INDEP VARIABLE	COEFFICIENT	PARTIAL CORR COEFF	TOLERANCE	R^2	$S_{y.x_1...x_n}$	F_s
CONSTANT	51.32977	-	-	-	-	-
COHRS	0.00795	-	-	0.8783	9.54	20.770
BAJOBS	0.01314	0.8861	0.8167	0.9738	4.72	19.090
TAPEHRS	-0.03434	0.4806	0.0303	0.9799	4.48	3.340
CPUHRT	3.95581	0.2569	0.7522	0.9812	4.74	28.350
LOGONS	0.04897	-0.2999	0.6009	0.9829	5.05	17.000
LNCARDIN	-0.04510	-0.4612	0.2532	0.9865	5.18	25.380
LOGHRS	-0.12651	-0.9525	0.0037	0.9987	1.93	19.560
OVERALL				0.9987	1.9	228.540

DEPENDENT VARIABLE: MEAN= 43.6965 STANDARD DEVIATION= 25.7818

Note: Independent variables are listed in order of entry.

APPENDIX 6

TABLE 20. MULTIPLE LINEAR REGRESSION RESULTS--AFIT/CREATE--
JUNE 1976 TO MARCH 1977: DEPENDENT VARIABLE--CPUHRT

INDEP VARIABLE	COEFFICIENT	PARTIAL CORR COEFF	TOLERANCE	R^2	$S_{y.x_1...x_n}$	F_s
CONSTANT	-13.00804	-	-	-	-	-
LOGHRS	0.03189	-	-	0.1320	2.14	56.730
LOGONS	-0.01236	-0.5291	0.0845	0.3750	1.94	41.890
LNCARDIN	0.01132	0.7117	0.3426	0.6915	1.47	86.030
TAPEHRS	0.00919	0.3798	0.3632	0.7360	1.49	6.020
COHRS	-0.00196	-0.5820	0.0493	0.8254	1.36	27.700
CPUHRB	0.23614	0.8516	0.0427	0.9461	0.87	28.350
BAJOBS	-0.00303	-0.8967	0.0564	0.9894	0.47	8.210
OVERALL				0.9894	0.5	26.810

DEPENDENT VARIABLE: MEAN= 7.4896 STANDARD DEVIATION= 2.1675

Note: Independent variables are listed in order of entry.

APPENDIX F

TABLE 21. MULTIPLE LINEAR REGRESSION RESULTS--AFIT/CREATE--
JUNE 1976 TO MARCH 1977: DEPENDENT VARIABLE--COHRS

INDEP VARIABLE	COEFFICIENT	PARTIAL CORR COEFF	TOLERANCE	R^2	$S_{y.x_1...x_n}$	F_s
CONSTANT	-6350.90198	-	-	-	-	-
TAPEHRS	4.85241	-	-	0.9365	929.50	12.830
CPUHRB	114.74998	0.7958	0.2532	0.9767	601.67	20.770
LOGHRS	15.59643	0.2362	0.9243	0.9780	631.49	42.620
LOGONS	-6.06305	-0.3706	0.0261	0.9810	642.50	38.240
LNCARDIN	5.46570	0.4032	0.2280	0.9840	657.35	33.030
CPUHRT	-474.91436	-0.8911	0.2612	0.9967	344.41	27.700
BAJOBS	-1.38526	-0.8347	0.0365	0.9990	232.19	4.600
OVERALL				0.9990	232.2	288.320

DEPENDENT VARIABLE: MEAN= 5026.9639 STANDARD DEVIATION= 3478.8037

Note: Independent variables are listed in order of entry.

APPENDIX F

TABLE 22. MULTIPLE LINEAR REGRESSION RESULTS--AFIT/CREATE--
JUNE 1976 TO MARCH 1977: DEPENDENT VARIABLE--TAPEHRS

INDEP VARIABLE	COEFFICIENT	PARTIAL CORR COEFF	TOLERANCE	R^2	$S_{y.x_1...x_n}$	F_s
CONSTANT	134.78050	-	-	-	-	-
COHRS	0.05557	-	-	0.9365	81.75	3.490
BAJOBS	-0.19782	-0.7232	0.8167	0.9697	60.36	5.430
CPUHRB	6.53576	0.4806	0.0262	0.9767	57.17	1.670
CPUHRT	-4.48513	-0.1774	0.7525	0.9774	61.64	0.160
LOGHRS*	-	-0.0295	0.5800	-	-	0.003
LOGONS*	-	0.0092	0.5469	-	-	0.001
LNCARDIN*	-	-0.0467	0.3503	-	-	0.009
OVERALL				0.9774	61.6	54.188

DEPENDENT VARIABLE: MEAN= 286.7736 STANDARD DEVIATION= 305.9801

Note: Independent variables are listed in order of entry.
*Variables not entered in the regression.

APPENDIX F

TABLE 23. MULTIPLE LINEAR REGRESSION RESULTS--AFIT/CREATE--
JUNE 1976 TO MARCH 1977: DEPENDENT VARIABLE--LOGHRS

INDEP VARIABLE	COEFFICIENT	PARTIAL CORR COEFF	TOLERANCE	R^2	$S_{y.x_1...x_n}$	F_s
CONSTANT	406.82677	-	-	-	-	-
LOGONS	0.38946	-	-	0.9155	113.80	1359.650
LNCARDIN	-0.34891	-0.2753	0.8529	0.9708	71.45	91.230
CPUHRT	30.28644	0.8496	0.9396	0.9904	44.22	56.730
CPUHRB	-7.17122	-0.3711	0.5983	0.9917	44.98	19.570
COHRS	0.06124	0.6687	0.0605	0.9954	37.39	42.620
TAPEHRS	-0.29071	-0.8609	0.0298	0.9988	21.97	7.560
BAJOBS	0.08747	0.8412	0.0378	0.9996	14.55	4.840
OVERALL				0.9996	14.5	827.010

DEPENDENT VARIABLE: MEAN= 1580.7330 STANDARD DEVIATION= 369.0832

Note: Independent variables are listed in order of entry.

APPENDIX F

TABLE 24. MULTIPLE LINEAR REGRESSION RESULTS--AFIT/CREATE--
JUNE 1976 TO MARCH 1977: DEPENDENT VARIABLE--LOGONS

INDEP VARIABLE	COEFFICIENT	PARTIAL CORR COEFF	TOLERANCE	R^2	$S_{y.x_1...x_n}$	F_s
CONSTANT	-1039.55217	-	-	-	-	-
LOGHRS	2.56390	-	-	0.9155	253.51	1359.650
LNCARDIN	0.89125	0.7875	0.9091	0.9679	167.04	64.340
CPUHRT	-77.24451	-0.8025	0.8633	0.9886	107.66	41.890
BAJOBS	-0.22065	0.3842	0.7021	0.9903	108.89	4.330
COHRS	-0.15674	-0.3343	0.1766	0.9913	114.73	38.240
CPUHRB	18.27283	0.8658	0.0133	0.9978	66.33	17.000
TAPEHRS	0.74495	0.8881	0.0223	0.9995	37.33	7.470
OVERALL				0.9995	37.3	623.350

DEPENDENT VARIABLE: MEAN= 3495.8000 STANDARD DEVIATION= 822.2084

Note: Independent variables are listed in order of entry.

APPENDIX F

TABLE 25. MULTIPLE LINEAR REGRESSION RESULTS--AFIT/CREATE--
JUNE 1976 TO MARCH 1977: DEPENDENT VARIABLE--LNCARDIN

INDEP VARIABLE	COEFFICIENT	PARTIAL CORR COEFF	TOLERANCE	R^2	$S_{y.x_1 \ldots x_n}$	F_s
CONSTANT	1164.89520	-	-	-	-	-
TAPEHRS	-0.81867	-	-	0.6050	197.97	7.100
LOGHRS	-2.80455	-0.3720	0.9892	0.6605	196.45	91.230
LOGONS	1.08819	0.5557	0.0453	0.7654	176.41	64.340
CPUHRT	86.36816	0.1122	0.0499	0.8976	127.64	86.030
COHRS	0.17251	0.5324	0.0455	0.9267	120.80	33.030
CPUHRB	-20.55123	-0.8627	0.0515	0.9812	70.54	25.380
BAJOBS	0.25901	0.8786	0.0484	0.9957	41.25	6.770
OVERALL				0.9957	41.3	66.520

DEPENDENT VARIABLE: MEAN= 1413.1700 STANDARD DEVIATION= 297.3576

Note: Independent variables are listed in order of entry.

APPENDIX G

CREATE: AGGREGATED DATA, GRAPHS AND RESULTS-- JANUARY 1974 TO MARCH 1977

APPENDIX G

TABLE 26. AGGREGATED DATA FOR TOTAL CREATE USAGE--
JANUARY 1974 TO MARCH 1977

MONTH	BAJOBS	CPUHRB	CPUHRT	COHRS	TAPEHRS	LOGHRS	LOGONS
JAN74	12064	385	72	43750	2003	8371	17436
FEB74	9647	392	98	38044	1888	8581	15749
MAR74	10842	429	124	37458	2066	10039	18277
APR74	11147	349	122	33771	2008	9299	18436
MAY74	13466	450	146	42507	2136	9535	18376
JUN74	11683	449	138	42522	2118	9737	17994
JUL74	13876	549	193	53051	2540	9674	18386
AUG74	13930	465	105	38321	2127	10002	21375
SEP74	14145	506	122	44702	2243	9776	20472
OCT74	12180	387	106	33537	1826	8932	18640
NOV74	13201	418	105	34643	2169	8678	17993
DEC74	10366	402	90	32399	1851	6970	14682
JAN75	11870	359	157	35519	2289	9472	18497
FEB75	11822	311	187	32199	1807	8930	17224
MAR75	14652	535	132	45094	2681	9738	20487
APR75	16108	572	134	48608	3430	9396	19620
MAY75	13090	404	84	35268	2045	8108	17577
JUN75	12692	562	122	48553	3198	9490	20632
JUL75	13984	659	127	43366	3077	10579	22625
AUG75	11780	442	129	22915	1771	8797	18599
SEP75	13000	500	130	24170	2376	10468	21361
OCT75	11988	396	113	20841	1932	8528	17178
NOV75	11125	359	99	20660	1761	7835	15854
DEC75	10421	281	80	22946	1244	6430	12856
JAN76	12121	298	108	24776	1151	8466	17058
FEB76	14370	311	85	27262	1243	8714	17630
MAR76	16142	483	94	37971	2004	9336	18933
APR76	13921	496	97	39433	1870	9375	18264
MAY76	14492	587	94	44471	1918	8701	16263
JUN76	12628	355	100	31414	1311	8183	15632
JUL76	7374	252	67	19876	1031	3542	7319
AUG76	14509	402	80	37224	1576	7849	16682
SEP76	12468	421	101	39420	1603	8669	18221
OCT76	11519	343	68	31949	1392	6280	13438
NOV76	13498	421	87	39235	2045	7322	15779
DEC76	11465	428	76	34645	1684	6261	12466
JAN77	10069	414	60	34675	1936	5890	12213
FEB77	11461	448	66	39734	2187	6670	13514
MAR77	15629	566	81	52102	3216	8377	16801

Note: Data have been truncated to nearest whole number.

APPENDIX G

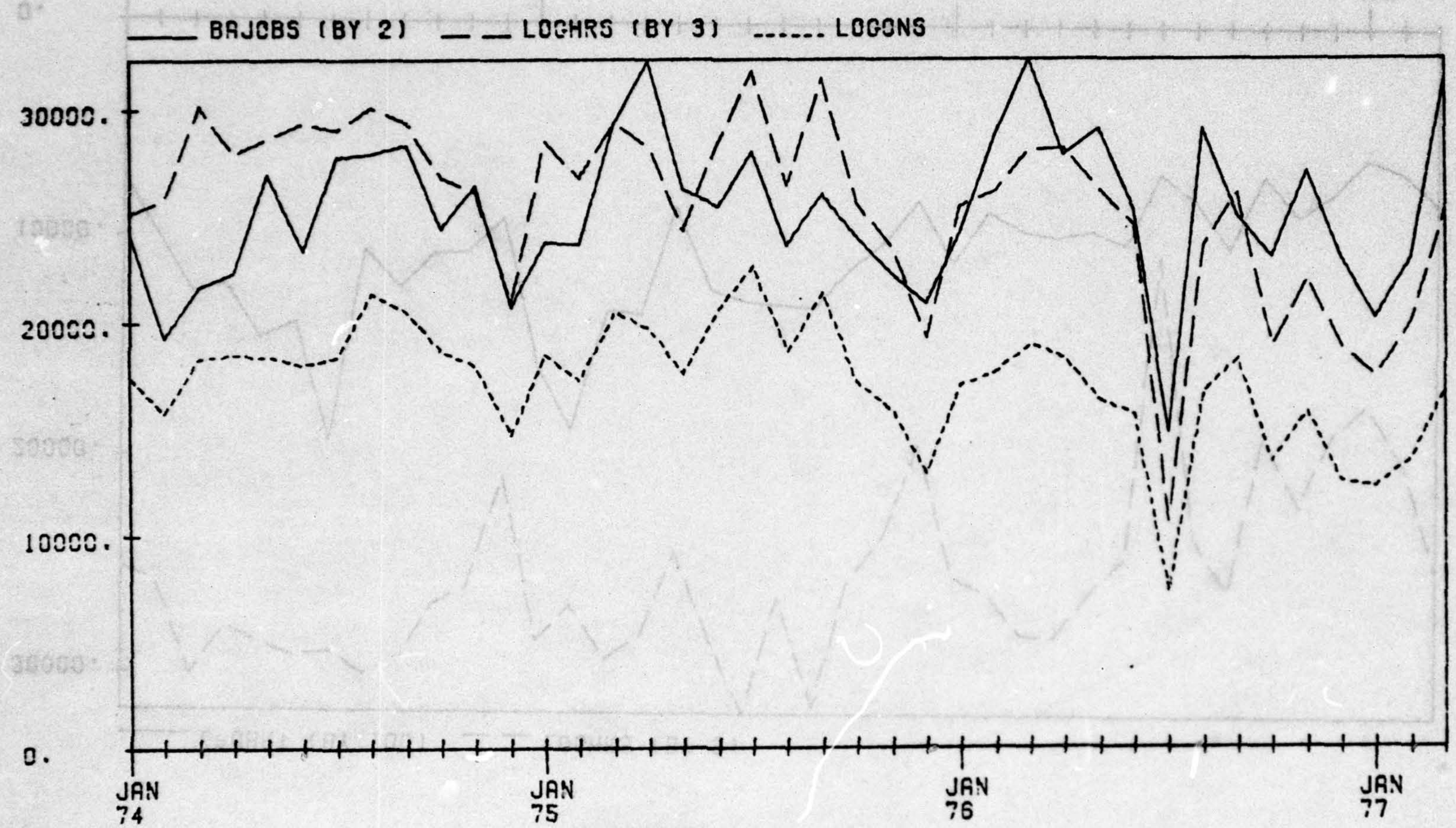


Figure 10. GRAPH: BAJOBS, LOGHRS AND LOGONS--CREATE

APPENDIX G

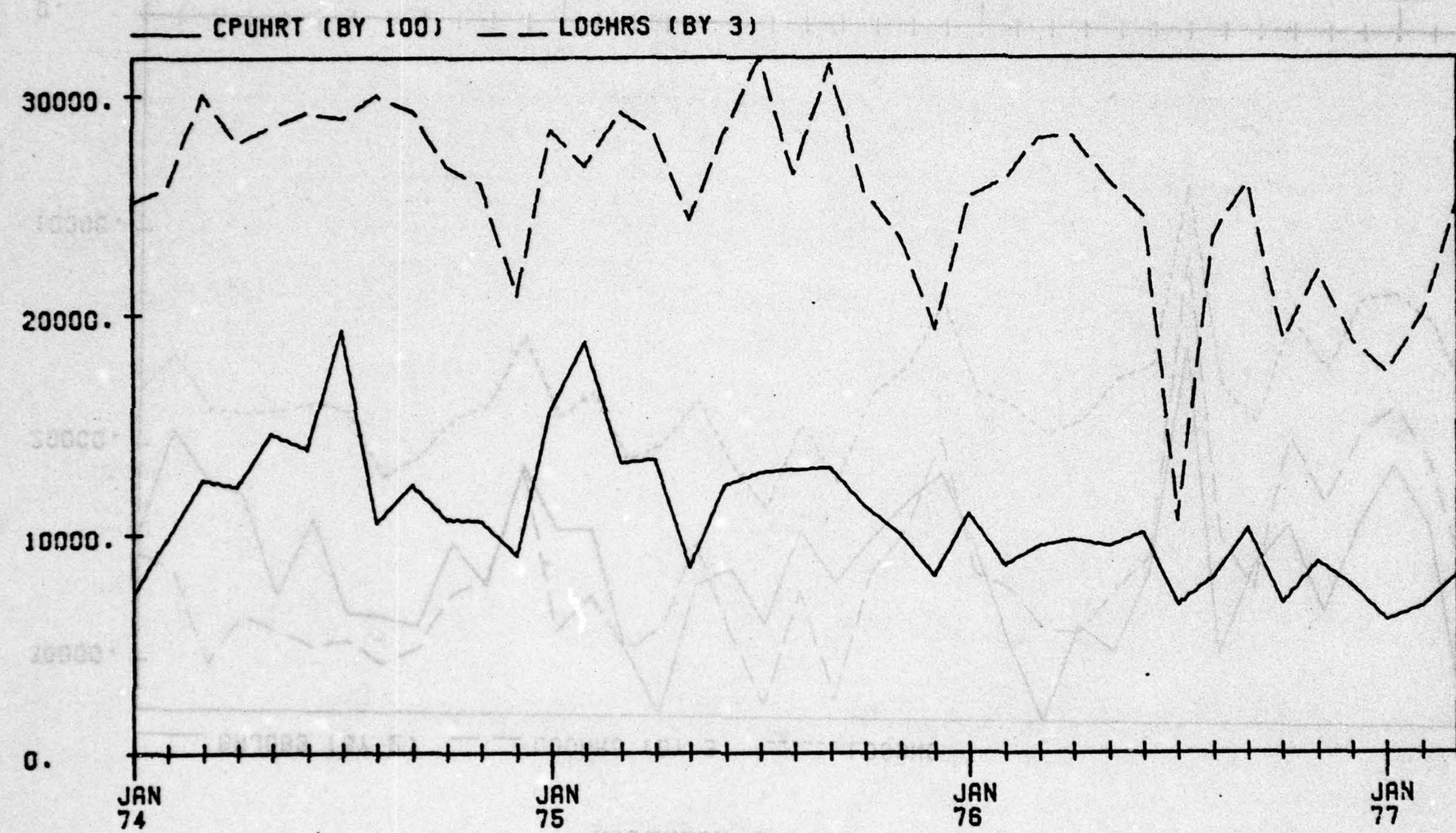


Figure 11. GRAPH: CPUHRT AND LOGHRS--CREATE

APPENDIX G

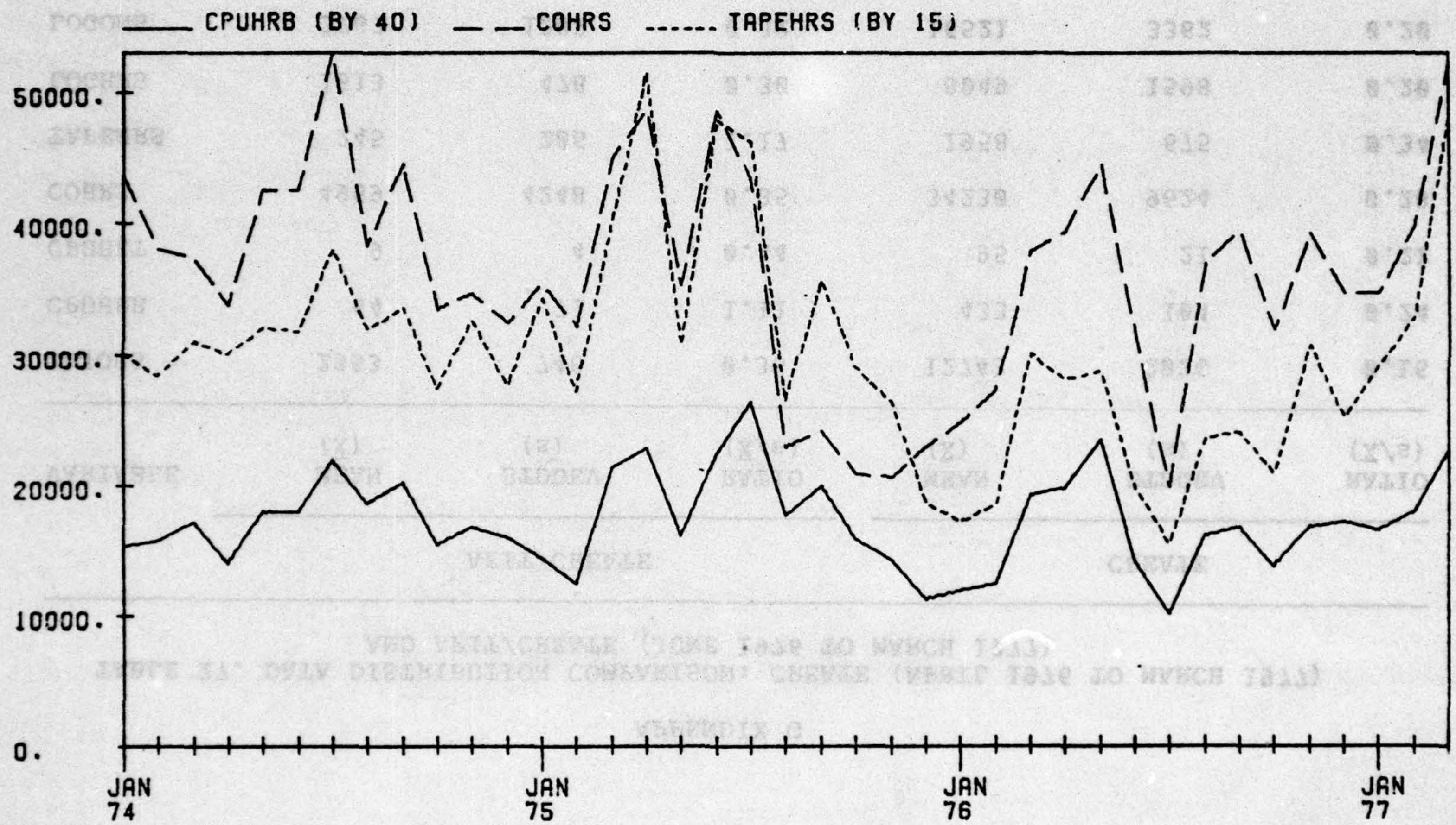


Figure 12. GRAPH: CPUHRB, TAPEHRS AND COHRS--CREATE

APPENDIX G

TABLE 27. DATA DISTRIBUTION COMPARISON: CREATE (APRIL 1976 TO MARCH 1977) AND AFIT/CREATE (JUNE 1976 TO MARCH 1977)

VARIABLE	AFIT/CREATE MEAN ($\overline{X}$)	AFIT/CREATE STDDEV (s)	AFIT/CREATE RATIO ($\overline{X}$/s)	CREATE MEAN ($\overline{X}$)	CREATE STDDEV (s)	CREATE RATIO ($\overline{X}$/s)
BAJOBS	2053	748	0.36	12743	2026	0.16
CPUHRB	64	71	1.11	433	104	0.24
CPUHRT	9	4	0.44	95	21	0.22
COHRS	4969	4248	0.85	34230	9624	0.28
TAPEHRS	245	286	1.17	1958	675	0.34
LOGHRS	1613	478	0.30	8049	1598	0.20
LOGONS	3787	1202	0.32	16521	3362	0.20

APPENDIX G

TABLE 28. PEARSON'S BIVARIATE CORRELATION COEFFICIENTS AND LEVEL OF SIGNIFICENCE--CREATE--APRIL 1976 TO MARCH 1977

	CPUHRB	CPUHRT	COHRS	TAPEHRS	LOGHRS	LOGONS
BAJOBS	.7736 .002	.5801 .024	.8545 .001	.6081 .019	.8867 .001	.8863 .001
CPUHRB		.3488 .133	.9252 .001	.7765 .001	.7126 .005	.6469 .011
CPUHRT			.3314 .146	-.0162 .480	.8005 .001	.7492 .003
COHRS				.8751 .001	.7439 .003	.7446 .003
TAPEHRS					.4045 .096	.4091 .093
LOGHRS						.9751 .001

APPENDIX G

TABLE 29. MULTIPLE LINEAR REGRESSION RESULTS: CREATE

VARIABLE	JAN74-MAR77 R^2	JAN74-MAR77 STANDARD ERROR	APR74-MAR77 R^2	APR74-MAR77 STANDARD ERROR	APR76-MAR77 R^2	APR76-MAR77 STANDARD ERROR
BAJOBS	.60526	1253	.80486	1041	.88103	1154
CPUHRB	.77603	47	.88032	40	.95921	27
CPUHRT	.58454	21	.80201	11	.81315	8
COHRS	.64747	5604	.82561	4678	.98595	1251
TAPEHRS	.7789	283	.83366	320	.92636	222
LOGHRS	.94560	368	.97849	273	.98790	266
LOGONS	.94343	762	.97159	659	.98607	545

APPENDIX G

TABLE 30. TCAST ANALYSIS--CREATE--APRIL 1975 TO MARCH 1977--LEAD TIME= 12

VARIABLE	CYCLE TCAST	CYCLE USED	CYCLIC ERROR	SMOOTHING TYPE	SMOOTHING ALPHA	MAD	TREND	MEAN	STDDEV
BAJOBS	1	1	289537	2	0.100	1942	13375-50.557t*	2743	2026
CPUHRB	1	1	547	1	0.150	132	470-2.921t	433	104
CPUHRT	5	5	18	2	0.150	70	125-2.39t	95	22
COHRS	1	1	597110	1	0.150	1085	31369+228.848t	4369	9624
TAPEHRS	1	1	17395	2	0.150	838	2297-27.116t	1958	675
LOGHRS	7	7	103606	2	0.150	850	9645-127.702t	8049	1598
LOGONS	7	7	387340	2	0.150	1733	20107-286.894t	6521	3361

*Time t is the number of months from April 1975.

SELECTED BIBLIOGRAPHY

A. REFERENCES CITED

1. Anderson, Captain Thomas E., USAF, and Gerald A. Purnell. "A Model to Forecast Computer Usage by the Air Force Institute of Technology." Unpublished master's thesis. SLSR 21-76B, AFIT/SL, Wright-Patterson AFB, OH, 1976. ADA032461.

2. Brown, Goodell Robert. Smoothing, Forecasting, and Prediction of Discrete Time Series. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1963.

3. Honeywell Information Systems. Time Series Forecasting. Waltham, Massachusettes, 1974.

4. Hunt, Earl, George Diehr and David Garnatz. "Who Are the Users?--An Analysis of Computer Use in a University Computer Center," AFIPS Conference Proceedings, Vol. 38, Spring Joint Computer Conference, 1971, pp. 231-238.

5. Jones, Lieutenant Colonel Eugene E., USAF. Assistant Professor of Computer Science, Department of Quantitative Management, School of Systems and Logistics, Air Force Institute of Technology, Wright-Patterson AFB, OH. Intermittent personal interviews from October 1976 to January 1977.

6. Lockett, Jo Ann. "Computer Performance Analysis In Mixed On-Line/Batch Workloads," AFIPS Conference Proceedings, Vol. 47, National Computer Conference, 1974, pp. 671-676.

7. Nie, Norman H. and others. Statistical Package for the Social Sciences, Second Edition. New York: McGraw-Hill Inc., 1975.

8. Reed, Maurice. Air Force Logistics Command, Deputy Chief of Staff Plans and Programs, Management Science Office (AFLC/XRS), Wright-Patterson AFB, OH. Personal Interview, 13 January 1977.

9. United States Air Force Institute of Technology. Catalogue: 1976-1978. Wright-Patterson AFB, OH, 1976.

10. United States Air Force Institute of Technology. Minutes of meeting of AFIT/AFLC CREATE Users Group, Wright-Patterson AFB, OH, 17 November 1975.

11. United States Air Force Logistics Command. "CREATE Management," AFLC Regulation 400-25. Wright-Patterson AFB, OH, 28 April 1972.

12. Watson, R. A. Measurement and Analysis of Computer System Performance: Applications of Accounting Data. The RAND Corporation, R-573-NASA/PR, September 1970.

B. RELATED SOURCES

Bedworth, David D. Industrial Systems--Planning, Analysis, Control. New York: The Ronald Press Co., 1973.

Bell, T. E. Computer Performance Analysis: Measurement Objectives and Tools. The Rand Corporation, R-584-NASA/PR, February 1971. AD721441.

Boehm, B. W. Computer Systems Analysis Methodology: Studies in Measuring, Evaluating, and Simulating Computer Systems. The Rand Corporation, R-520-NASA, September 1970.

Box, George E. P. and Gwilym M. Jenkins. Time Series Analysis and Forecasting Control. San Francisco, California: Holden-Day, Inc., 1970.